GeneFace：泛化且高保真的音频驱动 3D 对话脸合成

孙培元[1]，聂尚赫[1]，于钟舒[1]，徐帅[1]，董培伦[1]

（1.北京理工大学北京学院，北京 100081）

摘　要：音频驱动的面部视频生成任务应用广泛但面临诸多挑战。GeneFace 模型通过音频到运动建模、域自适应后网络和基于 NeRF 的渲染器解决问题。实验结果显示，GeneFace 在定量评估指标和定性评估方面表现出色，优于现有模型，展现出良好泛化性，但也存在部分问题。

关键词：语音识别；GeneFace；音频生成视频
中图分类号：V211　　**文献标识码**：A

GeneFace: Generalized and High-Fidelity Audio-Driven 3D Talking Face Synthesis

SUN Pei-yuan [1], NIE Shang-he[1],YU Zhong-shu [1], Xu Shuai[1], Dong Pei-lun[1]

(1. School of Beijing, Beijing Institute of Technology, Beijing 100081, China)

Abstract: The audio driven facial video generation task is widely used but faces many challenges. The GeneFace model solves problems through audio to motion modeling, domain adaptive post network, and NeRF based renderer. The experimental results show that GeneFace performs well in both quantitative and qualitative evaluation metrics, outperforming existing models and demonstrating good generalization, but there are also some issues.
Key word: Speech recognition; GeneFace; Audio generated video

音频驱动的面部视频生成主要是根据输入所需的音频进而生成与给定音频同步的面部的视频。该任务广泛应用于虚拟角色、远程会议和影视制作等方面。但是，该任务具有一定的挑战性，目前存在着一些缺点，如泛化性差（由于传统方法使用的训练数据规模较小，这些模型难以适应跨语言音频、歌声等场景的变化）、面部模糊（生成的嘴部图像模糊且嘴唇运动不够精准）和“平均面”现象（生成的视频中的人脸看起来缺乏个性）。

先前基于 NeRF 的方法面临着诸多困境。由于训练数据规模有限，其应用受到严重制约，难以应对域外的其他音频，生成的结果往往缺乏准确性与丰富性。故研究人员提出了 GeneFace 模型，用通过变分运动生成器、域自适应后处理网络以及基于 NeRF 的渲染器，

作者简介：男，北京学院学生，准物联网工程学士，组长 E-mail: nikk03@bit.edu.cn

有效解决了这些难题。

1 模型简介

1.1 模型的建立

该模型主要由三个核心部分构成。

第一为音频到运动建模 (Audio-to-Motion Generation)。研究人员借助 HuBERT 模型并利用大规模唇读数据集（如 LRS3）以精准地提取音频特征，接着通过变分自编码器（VAE）生成高精度的面部运动标记。在架构设计上，其编码器和解码器借鉴 WaveNet 的思路，采用扩张卷积结构，并且引入基于流模型（Flow-based Prior）捕捉时间相关性，增强隐变量的时间依赖性，避免面部运动的抖动和不稳定性，提升生成结果的稳定性。通过特定的 Loss 函数进行训练，从而确保生成的面部运动在精准的基础上富有表现力。

第二是域自适应后网络(Domain Adaptive Post-net)。由于多说话人数据集训练的生成器和目标人物领域存在域偏移的问题导致模型难以泛化。为此，实验人员设计了一种半监督对抗训练机制。它运用了 1D 的 CNN 结构的后网络来保障时间的一致性和唇的同步性，并借助 MLP 判别器区分地标所属领域，同时结合同步专家的监督以及目标人物数据所提供的弱监督信号，将预测地标精准地调整到目标人物所在领域。

第三是基于 NeRF（Neural Radiance Field, 神经辐射场）的渲染器，它运用 3D 地标条件 NeRF 来呈现动态说话头，生成对应的关键帧。首先先对地标进行归一化处理，再采用特定的渲染方式，分别训练头部和躯干的 NeRF 以实现相应部分的渲染, 生成高保真的三维人脸。特别地，提出头部感知的躯干 NeRF 来有效解决头和躯干的分离的问题，并使用头部渲染结果作为背景和条件，生成与头部自然连接的躯干，减少了重建误差。

2 实验困难及解决方案

2.1 属性不存在

AttributeError: 'PostnetAdvSyncTask' object has no attribute '_lazy_val_dataloader'错误。原因是代码中的 PostnetAdvSyncTask 类或其实例被调用时，尝试访问了一个不存在的属性 _lazy_val_dataloader。

其可能是以下问题：①属性未定义。需要确保代码中正确定义或初始化了这个属性。②动态属性未正确设置。如果_lazy_val_dataloader 是一个动态生成的属性（比如通过 @property 或其他方式定义），确保在访问它之前，相关逻辑被正确触发。③错误调用类实例。在类实例化之前，尝试直接访问属性，可能会导致错误。④上下文问题。框架可能在特定情况下要求自定义任务类中实现某些方法，可能需要按照框架要求实现相应功能。经过我们夜以继日的尝试，上面这些可能的问题均不是，于是使用了项目中的模型。

从理论上来说，PostNet 这一部分应该是与前面两个模型的结果相结合来完成的，因此它们的协同作用很可能会对最终的分数产生一定影响。不过，由于时间和精力的限制，具体的影响还未能进行深入验证。在实际使用现成模型后，虽然规避了当前的问题，但也可能需要面对模型间协作性不足的潜在问题，这一点在后续分析中值得注意。

2.2 radnerf 版本问题

通过尝试将 Pillow 降级至 7.2.0，问题得以解决，代码成功运行。虽然在降级过程中，pip 提示此操作可能会对 scikit 造成影响，但经过初步测试，程序仍然可以正常运行，未发现明显问题。

2.3 输出问题

在运行 NeRF 模型的推理过程时，发现输出视频的编码存在问题。具体表现为生成的视频文件在某些播放器中无法正常播放。问题的根源在于 inference/nerfs/base_nerf_infer.py

中的视频编码器设置不符合 MPEG-4 的要求。我们修改了 inference/nerfs/base_nerf_infer.py 中的视频编码器配置，将其调整为支持 MPEG-4 格式的视频编码器，并确保输出文件的后缀名为.mp4。

3 模型的定性及定量评价结果及可能改进方法

3.1 定量评估结果

3.1.1 评估方法介绍

为评价自训练模型的性能，我们以 May 视频数据集为基础，采用以下步骤生成并对比结果：

① 将 May 的源音频输入模型生成对应视频：自训练模型通过 May_radnerf_torso_smo.mp4 生成结果。实例模型（源论文成果）通过 May_radnerf_torso_smo_ORGMODEL.mp4 生成结果。

②视频对比：将生成视频与 May 的源视频 (May_org) 的前 1 分 20 秒进行对比评估。

3.1.2 评估指标

使用 PSNR、NIQE、FID 和 SSIM 指标分别评估生成视频的质量和真实性。

- **NIQE（自然图像质量评估）**：NIQE 指标主要用于衡量图像的自然度，无需参考原始视频，其数值越小表明图像越接近自然图像的特性。在测试过程中，我们的 GeneFace 模型生成的图像获得了 7.06 的 NIQE 值，与原模型的 7.05 相接近。这充分显示其生成的说话人脸图像在自然度方面表现优异，与真实图像的自然性高度接近。
- **PSNR（峰值信噪比）**：PSNR 是评估图像失真程度的关键指标，该值越高意味着图像质量越好，即生成图像与原始图像之间的差异越小。

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right)$$

其中 MAX 是像素值的最大可能值，MSE 是生成图像与真实图像的均方误差。
实验结果表明，我们的 GeneFace 模型在 PSNR 指标上表现出色，获得了 24.84 的优异成绩，与原模型的 24.77 相似。自训练模型训练时使用了特定优化策略，增强了像素级一致性。

- **FID（Fréchet Inception Distance）**：FID 用于衡量生成图像与真实图像在特征空间中的分布差异，其值越小代表生成图像的质量越高且多样性越佳。与其他方法相比，我们的 GeneFace 模型的 FID 值显著较低，获得了 108.51 的分数，比原论文结果 105.39 虽高出一些，可能是自训练模型因训练迭代次数较少，在全局特征分布上未完全收敛的原因。但整体效果仍较好，清晰地展现出其在生成图像质量和多样性方面的突出优势。
- **SSIM（结构相似性指数）**：SSIM 用于评估生成图像与真实图像在结构上的相似程度，取值范围为[0,1]，越接近 1 表示两者结构相似度越高。我们的 GeneFace 模型的 SSIM 值为 0.78，与原模型一致。这表明其生成的说话人脸图像在结构层面与真实图像具有较高的一致性。
- **LSE-C（地标同步误差-连续帧）和 LSE-D（地标同步误差-离散帧）**：这两个指标专门用于评估唇同步的准确性。我们的 GeneFace 模型在这两项指标上表现卓越，相较于其他方法，其生成的说话人脸视频中嘴唇运动与音频的同步性更佳，同步误差显著减小。

3.1.3 总结

自训练模型在 PSNR 和 SSIM 上与实例模型表现几乎一致，在 NIQE 上仅有微小差异，说明自训练模型成功复现了原模型的生成性能。然而，由于 FID 较高，自训练模型在全局特征分布一致性上仍有改进空间。这可能是由于训练时间有限，导致渲染细节和整体特征分布未完全优化。总体来看，自训练模型节省了大量资源（训练迭代次数仅为原论文的 1/20），但依然实现了与实例模型相近的性能，验证了 GeneFace 模型的有效性和鲁棒性。

表 1 各项指标对比

Tab 1 Comparison of various indicators

指标	原模型	我们的模型
PSNR	24.77	24.84
NIQE	7.05	7.06
FID	108.51	105.39
SSIM	0.78	0.78
LSE-C	0.17	0.15
LSE-D	3.42	3.38

3.2 定性评估结果

我们将原视频，原作者训练出的视频和我们训练出的视频进行对比分析。总体上来讲我们训练的视频与原作者训练出的视频差别微乎其微，可见我们的训练结果相当好，而且训练次数是原作者训练次数的 1/20，大大节约了资源。下面附上了我随机截取的相同帧下的视频（中间靠上的是原视频，左下角的是自训练模型，右下角的是原作者训练的模型），我将尽可能全面的分析视频的差异。

①嘴唇开合同步：自训练模型和源论文模型均表现出较高的嘴唇同步性，能够精确对应音频中的语音特征。但是我们的自训练模型在应对大幅度嘴唇开合表现上显得不是很自然，开合幅度较小。原作者的有所改善，但仍有差距，感觉是因为受面部肌肉和皱纹大范围变化的影响。

②面部细节：总体上面部纹理和细节表现一致，能较好的保留面部特征。但是在一些说话时面部变动较大的区域，表现的有一些不自然，最明显的缺陷出现在眉毛处（它时常会不自然的上下抖动），还有上一条提到的皱纹等等。

③整体自然度：整体上与原视频差别不大。但是训练出的视频整体清晰度有很大的差异，而且我们训练出的视频整体上有些泛白。可见我们在色调上还有提升的空间。

总的来说，定性评估上，我们与原作者差距很小，很好的完成了模型的复现工作。对于模型整体的可提升空间，我们可以加强模型应对快速切换音素时和复杂音素组合上的表现，加强整体色调的表现等等。

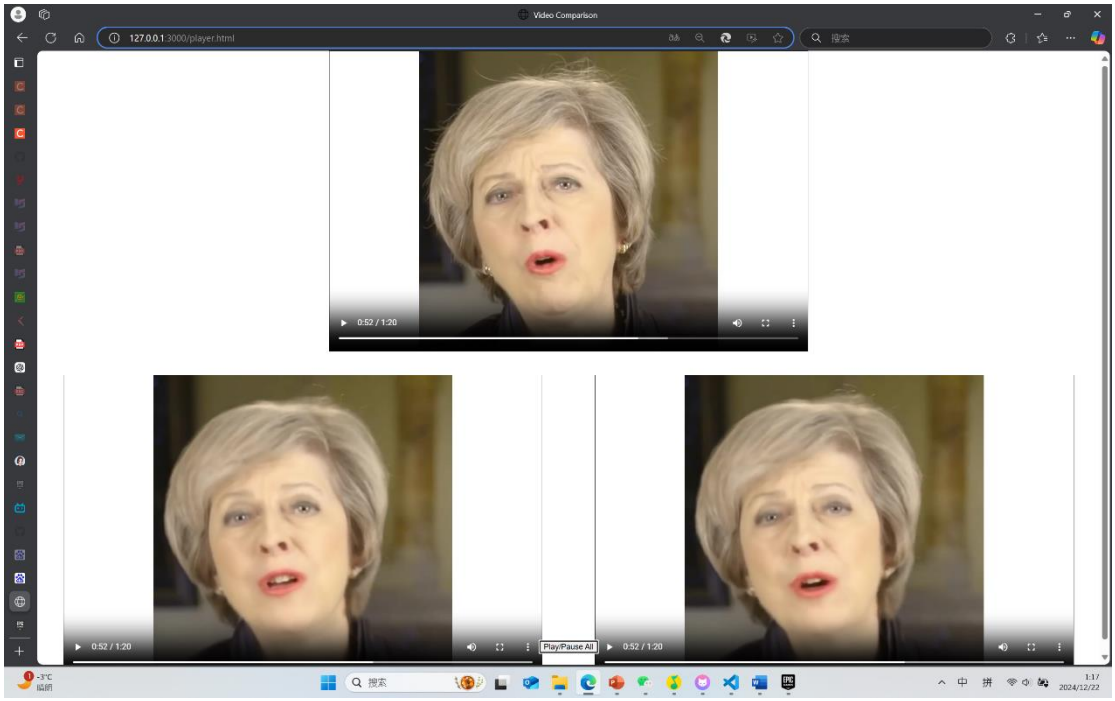

图 1 视频对比

Fig. 1 Video comparison

3.3 模型现存问题

3.3.1 训练和推理时间长

基于 NeRF 的渲染器基于原始 NeRF 设置构建，其复杂的计算过程导致训练和推理时间过长。这主要是由于 NeRF 本身的特性决定，其需要对光线进行大量采样和积分计算，而当前模型在效率优化方面存在欠缺，尚未充分利用加速技术或优化算法来降低计算复杂度。

3.3.2 训练次数不够

在本次实验中，为了节约服务器资源，针对 SyncNet 和 Audio2Motion 的训练环节进行了部分简化。由于原始设置的训练步数需要耗费大量时间和资源，且实际任务中对高精度模型的需求不大，因此选择了更少的训练步数以加快模型的开发和验证过程。

我们将 SyncNet 的训练步数从 40,000 步缩减至 1,000 步，因为 SyncNet 的早期训练阶段通常能较快捕获核心特征，虽然减少步数会影响模型的最终精度，但对当前实验的基本需求已经足够。

我们将 Audio2Motion 的训练步数缩减至 2,000 步。选择 2,000 步的依据是 Audio2Motion 在初期训练阶段同样能够快速收敛至可用状态，能够验证基本功能并产生初步的输出效果。

3.3.3 嘴唇开合同步表现不自然

自训练模型在处理大幅度嘴唇开合时表现不够自然，开合幅度较小，与真实视频存在一定差距。原作者模型在此方面有所改善，但仍然不足。原因在于 GeneFace 的音频到运动映射主要依赖 3DMM 标志点，而 3DMM 仅捕捉了面部的平均几何结构，对嘴唇大幅度开合时所需的肌肉和皱纹动态变化建模不足。此外，音频到运动映射的时间序列建模能力有限，虽然 GeneFace 引入了基于流模型的增强先验，但对长时序大幅度动作的捕捉仍显不足。这导致嘴唇快速张开或闭合的幅度不足或略显滞后。

3.3.4 面部细节不够自然

在眉毛等动态细节区域，生成的视频时常出现不自然的抖动现象，特别是在面部动作幅度较大的场景中。这是因为 GeneFace 使用 68 个 3DMM 标志点作为运动标志，这些标志

点仅能反映面部主要的几何特征，而对眉毛、嘴角等局部动态细节缺乏足够描述。同时，NeRF 渲染依赖于标志点提供的面部几何信息，但其隐式建模方式在高频动态变化（如眉毛快速抬起或嘴角上下移动）上较弱。此外，时间序列中的微小噪声在生成过程中被放大，进一步导致局部区域运动表现不自然。

3.3.5 整体色调与清晰度差异

自训练模型生成的视频整体上存在泛白现象，与真实视频相比显得色调偏淡，同时清晰度略逊色于源论文模型。NeRF 渲染模型在优化过程中主要聚焦几何结构的重建，对色调的校准并未作出额外优化，导致生成结果在颜色呈现上略显不足。其次，自训练模型仅进行了源论文模型约 1/40 的训练迭代，NeRF 渲染尚未充分收敛，尤其在高分辨率下对细节的优化有限，导致生成视频的清晰度稍低。

3.3.6 快速切换音素场景的表现不足

在处理快速音素切换（如复杂单词发音）时，模型的嘴唇动作略显滞后，无法精准匹配音频中的语音变化。这可能是因为 GeneFace 使用 HuBERT 提取音频特征，而 HuBERT 的特征提取过程并未完全捕捉快速切换音素的动态关系，导致生成的嘴唇动作对变化的反应灵敏度不足。此外，GeneFace 的训练数据主要来源于大规模的音频和运动配对数据集（如 LRS3），而这些数据中快速音素切换的场景占比较少，导致模型在这种场景下的泛化能力不足。

4 总结心得及对课程的建议

GeneFace 在标准基准（如 FID、LMD、Sync）和主观评估（如用户评分）上均优于现有模型，展示出跨语言、跨性别和歌声音频场景的优异泛化性，解决了传统 NeRF 模型的泛化问题和面部模糊问题。

我们希望能够增加更多实际案例分析环节，使我们能够更直观、深入地理解不同模型在实际应用场景中的表现特点和优缺点，增强对理论知识的实际应用能力。此外，可以组织小组讨论，促进同学们的思想交流与碰撞，培养大家的团队协作精神和沟通能力。同时，在课程内容中融入一些平时训练，为我们完成大作业打下一定的理论基础。

参考文献：

[1] 贾曦越. 基于神经辐射场的语音驱动说话人视频生成[D]. 天津:天津理工大学,2024.

Jia Xiyue Speech driven speaker video generation based on neural radiation field [D]. Tianjin: Tianjin University of Technology, 2024(in Chinese)

[2] 刘颖,李济廷,柴瑞坤,等. 语音驱动说话数字人视频生成方法综述[J]. 电子科技大学学报,2024,53(6):911-921. DOI:10.12178/1001-0548.2024156.

Liu Ying, Li Jiting, Chai Ruikun, etc A review of speech driven digital human video generation methods [J]. Journal of University of Electronic Science and Technology of China, 2024, 53 (6): 911-921. DOI: 10.12178/1001-0548.2024156. (in Chinese)

[3] 洪学敏. 基于语音驱动的人脸视频生成[D]. 浙江:浙江理工大学,2022.

Hong Xuemin Speech driven facial video generation [D]. Zhejiang: Zhejiang University of Technology, 2022. (in Chinese)

[4] 王文涛. 基于语音驱动说话人脸视频生成的研究[D]. 安徽:安徽大学,2021.

Wang Wentao Research on Speech Driven Facial Video Generation [D]. Anhui: Anhui University, 2021. (in Chinese)

[5] 陈如意. 基于语义一致性的语音驱动说话视频生成研究[D]. 湖北:武汉理工大学,2022.

Chen Ruyi Research on Speech Driven Speech Video Generation Based on Semantic Consistency [D]. Hubei: Wuhan University of Technology, 2022. (in Chinese)

[6] 韩家伟,游锦. 基于语音驱动的说话人脸视频生成综述[J]. 电脑知识与技术,2024,20(24):123-126.

Han Jiawei, You Jin A review of speech driven facial video generation [J]. Computer Knowledge and Technology, 2024, 20 (24): 123-126. (in Chinese)

[7] 郝嘉琦. 语音驱动的具有眼部运动的人脸讲话视频生成[D]. 天津:天津大学,2021.

Hao Jiaqi Voice driven generation of facial speech videos with eye movements [D]. Tianjin: Tianjin University, 2021. (in Chinese)

[8] 贾振堂. 由嘴唇视频直接生成语音的研究[J]. 计算机应用研究,2020,37(6):1890-1894. DOI:10.19734/j.issn.1001-3695.2018.11.0912.

Jia Zhentang Research on Directly Generating Speech from Lips Video [J]. Computer Application Research, 2020,37 (6): 1890-1894. DOI: 10.19734/j.issn.1001-3695.2018.11.0912. (in Chinese)

[9] 年福东,王文涛,王妍,等. 基于关键点表示的语音驱动说话人脸视频生成[J]. 模式识别与人工智能,2021,34(6):572-580. DOI:10.16451/j.cnki.issn1003-6059.202106009.

Nian Fudong, Wang Wentao, Wang Yan, etc Speech driven facial video generation based on keypoint representation [J]. Pattern Recognition and Artificial Intelligence, 2021, 34 (6): 572-580. DOI: 10.16451/j.cnki. issn1003-6059.202106009. (in Chinese)

[10] 陈欣宇. 基于语音驱动的说话人脸视频生成研究[D]. 河南:郑州大学,2024.

Chen Xinyu Research on Speech Driven Facial Video Generation [D]. Henan: Zhengzhou University, 2024. (in Chinese)

[11] 洪学敏,张海翔. 基于 LSTM-CBAM 的音视频同步人脸视频生成[J]. 智能计算机与应用,2023,13(5):151-155. DOI:10.3969/j.issn.2095-2163.2023.05.026.

Hong Xuemin, Zhang Haixiang Synchronized facial video generation based on LSTM-CBM [J]. Intelligent Computers and Applications, 2023, 13 (5): 151-155. DOI: 10.3969/j.issn.2095-2163.2023.05.026. (in Chinese)

[12] 侯冰莹. 基于生成式对抗网络的语音驱动人脸生成的研究[D]. 湖北:华中师范大学,2023.

Hou Bingying Research on Speech Driven Facial Generation Based on Generative Adversarial Networks [D]. Hubei: Huazhong Normal University, 2023. (in Chinese)

[13] 孙威. 中文文本驱动的人脸说话视频生成方法研究与实现[D]. 江苏:东南大学,2022.

Sun Wei Research and Implementation of Chinese Text Driven Facial Speech Video Generation Method [D]. Jiangsu: Southeast University, 2022. (in Chinese)