

单次自由视角神经说话头部合成

陈松宇, 梁思源, 王子鸣, 郭宇航

摘要: 本实验为“单次自由视角神经说话头部合成”(One-Shot Free-View Neural Talking Head Synthesis) 的项目复现, 旨在为视频会议等应用场景提供高质量、动态、自由视角的人物面部合成。通过利用源图像和驱动视频, 本方法能够在不同视角下合成出逼真的人物头部动画, 支持自由视角的动态效果, 包括偏航、俯仰和滚转等。项目复现完成后, 笔者还进行了多角度评估, 并提出了一些创新。项目复现使用了 PyTorch 1.7 和 Python 3.7 的环境, 支持多 GPU 训练, 提供了多个预训练模型, 并且在生成过程中优化了旋转矩阵以替代传统的雅可比矩阵。通过该方法, 可以生成高质量的虚拟人物面部动画, 具有较强的适应性和鲁棒性, 能够满足各种视频合成需求。

关键词: 神经头部动作合成; 自由视角; 面部动画; 深度学习

One-Shot Free-View Neural Talking Head Synthesis

Chen Songyu, Liang Siyuan, Wang Ziming, Guo Yuhang

Abstract: This experiment is a reproduction of the "One-Shot Free-View Neural Talking Head Synthesis" project, aiming to provide high-quality, dynamic, and free-viewpoint facial synthesis for applications such as video conferencing. By utilizing source images and driving videos, this method can synthesize realistic head animations from different viewpoints, supporting dynamic effects from free viewpoints, including yaw, pitch, and roll. After completing the project reproduction, the author also conducted multi-angle evaluations and proposed some innovations. The project reproduction used a PyTorch 1.7 and Python 3.7 environment, supported multi-GPU training, provided several pre-trained models, and optimized the rotation matrix during the generation process to replace the traditional Jacobian matrix. This method enables the generation of high-quality virtual human facial animations, with strong adaptability and robustness, capable of meeting various video synthesis requirements.

Key word: Neural Head Motion Synthesis; Free-Viewpoint; Facial Animation; Deep Learning

1 组内分工及组内评价

1.1 组长对组员贡献评价 (0-1.0)

组员姓名	个人贡献度
梁思源	1.0

王子鸣 1.0

郭宇航 1.0

1.2 组员对组长评价（0-1.0）

组长个人贡献度为各位组员对组长评价的平均值

组员姓名	对组长个人贡献度的评价
梁思源	1.0
王子鸣	1.0
郭宇航	1.0

2 实验报告

2.1 模型简介（0.05*个人贡献度*项目分）

2.1.1 项目背景

在当今时代背景下，全球图像识别软件市场呈现出蓬勃发展的态势，其规模正以迅猛之势持续扩张，预计在未来数年将以颇为可观的复合年增长率保持上扬态势。在此趋势推动下，图像识别领域的人工智能模型顺势而生，并凭借其卓越的性能和广泛的适用性，迅速渗透至各个行业，成为推动产业发展与变革的关键力量，在众多领域发挥着不可或缺的作用，为各行业的数字化转型和智能化升级注入了强劲动力。

2.1.2 项目介绍

本项目以“用于视频会议的一次性自由视角神经说话头部合成”这一前沿论文为理论根基，聚焦于运用神经网络技术构建高度逼真的说话者头部动态图像，旨在为视频会议场景中的虚拟人物创建与优化提供创新性解决方案。其核心理念在于借助先进的深度学习模型，凭借给定的特定输入数据，精心打造具备自由视角切换功能且动态效果极为逼真的说话者头像，从而显著优化视频会议参与者的交互体验，使沟通更加生动、自然且富有沉浸感。

从形式化的角度阐释，本项目的实施过程具体表现为：将一张目标头像（静态图片）以及一个驱动视频作为原始素材输入系统，通过复杂的算法处理与模型运算，最终生成以驱动视频中的表情、嘴型、头部运动等特征为参照，目标头像能够相应呈现动态变化的全新视频。

简而言之，该项目实现了一种独特的视觉效果转化，即利用他人的动作和表情数据，巧妙驱动一张静态图片，使之鲜活起来，生成一段在视觉上呈现出“此图片中的人物正在自然说话和移动”效果的视频，为视频会议领域的虚拟形象呈现开拓了新的可能性。

2.1.3 技术运用

该项目包含主要的四个神经网络模型：

生成器模型：

1. 初始化：一个 DenseMotionwork 网络，其内部使用了 Hourglass 这一网络架构：通过多层卷积的下采样（Encoder）提取特征和上采样（Decoder）恢复分辨率的对称式结构。下采样阶段降低了图像分辨率，通过逐步的下采样块，指数式地增加通道数，获得图片的高级抽象语义信息。上采样阶段恢复特征图大小，将上采样后的特征图与下采样（编码器）的跳跃连接特征图拼接，使得网络结构能包含全局信息。

2. 前向传播：构建一个初始化卷积块，使用下采样块压缩分辨率，接着通过一个卷积块限制通道数，再展开为 3D 特征图并进行处理，通过 DenseMotionwork 网络估计密集运动

和计算变形与遮罩层进行处理，最后通过上采样解码。

判别器模型 (MultiScaleDiscriminator):

主体类似 Pix2Pix，通过对输入图像进行下采样提取特征，再进行与生成器类似的压缩分辨率增大通道数，最后通过卷积层输出多尺度特征图。在此基础上将单个判别器更具不同尺度进行组合，获得不同尺度下的特征信息（虽然最后只用到了一个尺度）

生成器和判断器构成了生成对抗网络 (GAN)，二者相互竞争，增强生成器生成视频能力的同时也不断提高判别器的分辨能力

关键点检测模型 (KPDetector):

1. 生成一个 Hourglass 架构模型，通过连续下采样压缩空间并限制通道数，再通过卷积层变换通道数，重塑特征图维度后调整形状，最后使用上采样块恢复空间分辨率。

2. 建立 3d 卷积层，根据预测的关键点输出热图，通过在关键点周围生成高斯分布表示关键点的概率（高斯热图回归），将特征图关键点位置转换为概率分布，进而增强平滑性和鲁棒性，并在归一化时调用 softmax 函数控制热图的平滑度。

3. 估计雅可比矩阵，再使用一层卷积层从特征图中生成雅可比矩阵，将矩阵和对应热图点乘计算加权平均，获得关键点周围的雅可比矩阵，表示局部变换。

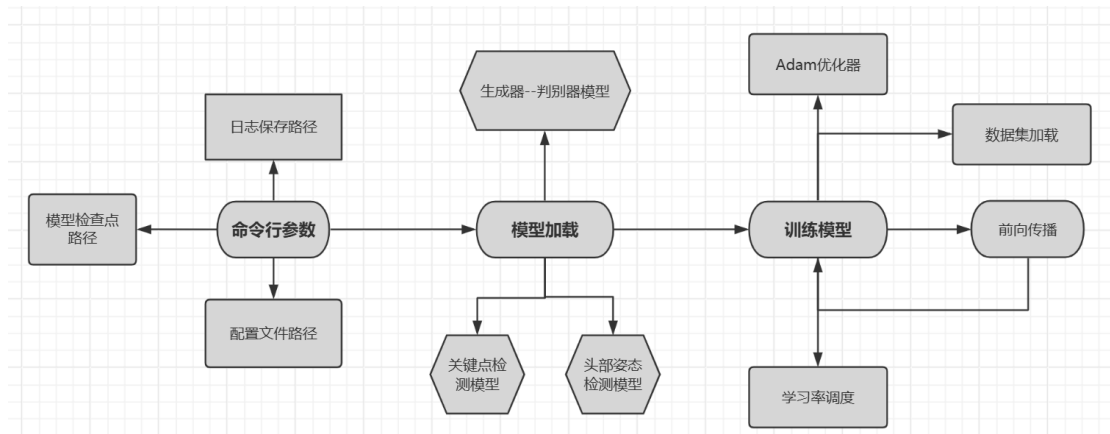
4. 前向传播：在缩放因子不为 1 时首先使用高斯卷积核进行下采样，再通过 Hourglass 架构的关键点捕获多尺度的特征，生成特征图后再通过 3d 卷积层整形后获得热图，归一化并重塑后再从热图中提取关键点的均值，最后按需计算雅可比矩阵。

姿态估计模型 (HEEstimator):

通过一系列的卷积层提取输入图像特征，再通过残差网络提炼特征，最后通过全连接层估计姿态。

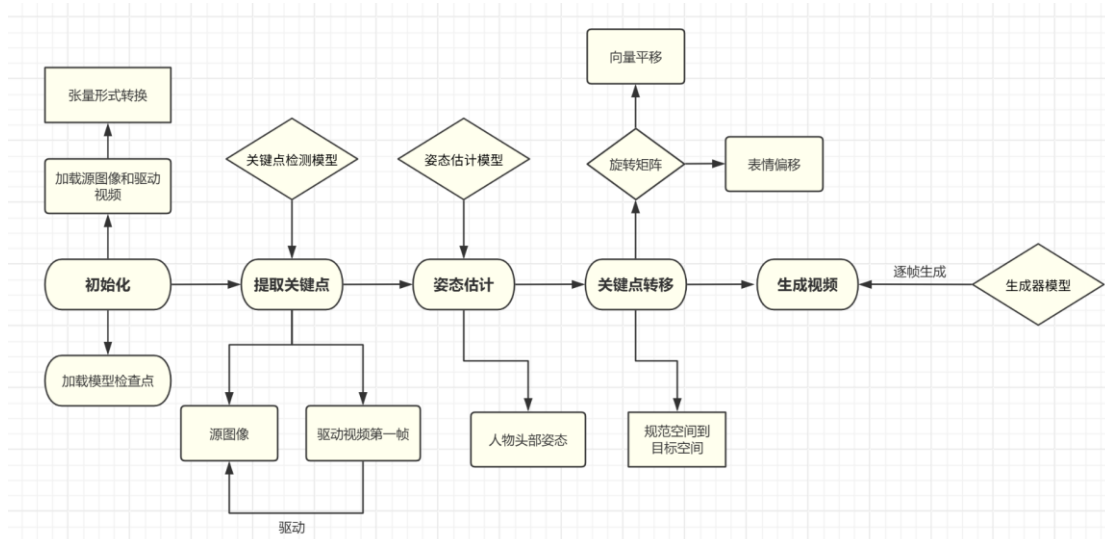
训练脚本 (train.py):

首先定义命令行参数，包括配置文件路径，日志保存路径，模型检查点路径等。接着设置了生成器模型、判别器模型、关键点检测模型和头部姿态检测模型，最后加载数据集进行训练。



测试脚本 (demo.py):

首先初始化，需要加载源图像和驱动视频，同时加载模型检查点，将源图像和驱动视频转换为张量形式。接着使用关键点检测模型检测源图像和驱动视频第一帧的关键点，从而实现基于驱动视频的关键点运动驱动源图像的动画。然后使用姿态估计模型获取人物的头部姿态，通过旋转矩阵对关键点旋转后再接平移和表情偏移，将关键点从规范空间姿态转换到目标姿态（可选自由视角），从而模拟人物的三维图像转动而不是仅限于二维。再提取驱动视频每一帧的关键点，同样经过姿态转换到目标空间，最后再进行标准化，通过生成器模型即可生成对应视频。



一次性自由视角（One-shot Free-viewpoint）技术不需要大量的训练数据或者视角数据，系统可以从少量的样本或者一个视角中生成不同角度的头像视图。与传统的视频渲染方式相比，生成新的视角的成本较低。

神经说话头部合成（Neural Head Synthesis for Talking Faces）技术使用神经网络（如卷积神经网络、生成对抗网络等）来模拟人类面部表情和动作，生成自然的嘴型、面部表情变化等，尤其是在说话的过程中。模型可能会结合音频和已有的图像数据来合成嘴唇运动和面部表情。

2.2 实验困难及解决方案（0.1*个人贡献度*项目分）

在本项目的推进过程中，初期便遭遇了棘手的难题：项目未提供明确的 requirements，这无疑为后续的环境配置工作带来了极大的挑战。在面对给定的代码时，众多库的具体版本信息缺失，这使得在尝试确定合适版本的过程中困难重重。通过向 AI 求助，虽然获得了几个可行的版本建议，但在实际测试中发现，对于单个或多个脚本能够正常运行的库，彼此之间存在严重的兼容性问题，这就要求在众多版本中仔细权衡、反复调试，以寻找各库版本之间的平衡状态，确保整个项目环境的稳定运行。

此外，由于此电脑在初始注册阶段，未曾充分考虑编码相关事宜，将用户名设置为中文，而部分较为陈旧的包在处理中文路径时存在明显缺陷，无法进行正确编码，进而引发了大量的报错信息，直接导致环境编译工作受阻失败。例如，在尝试下载 python-opencv 库并构建 wheel 文件时，频繁出现异常奇怪的错误提示。为解决这一问题，先是尝试使用清华源进行下载，同时发现部分包的临时文件默认存储在 C 盘用户名文件夹下（因用户名是中文而无法被识别），从而导致报错。通过深入研究，在原脚本中对临时文件的存储位置进行了修改，将其指定到一个英文路径的文件夹下，成功绕过了这一障碍，使问题得以顺利解决。

历经数小时的不懈尝试，最终成功完成了环境配置工作。值得一提的是，在项目前期由于未注意到文档中的裁剪脚本，自行编写了一个裁剪版本，而后又使用了文档中的脚本重新进行了裁剪操作。

然而，尽管已经克服了诸多困难，但编码问题依然如影随形。在运行脚本时，发现最终生成文件的名称路径存在异常，目前暂时采取手动设置路径的方式予以应对，并计划在封装 docker 阶段对这一问题进行全面优化改进，以确保整个流程的自动化与稳定性。

另外，关于生成的视频没有声音这一问题，经调查发现原项目本身输出的即为无声视频。为弥补这一不足，后续补充编写了部分代码，并独立开发了专门的脚本，用于将两个视频的

声音进行拼接，并在此基础上开展了一系列音频同步相关的测试工作，例如针对 LSE-C、LSE-D 的唇形匹配测试，以进一步提升视频的完整性和质量。

2.3 模型的定性和定量评价结果及可能改进方法（0.3*个人贡献度*项目分）

2.3.1 定量评估结果

这里我们选择了 NIQE , PSNR, FID, SSIM, MS-SSIM, LSE-C, LSE-D, L1, LPIPS 等测评指标，并使用给定的测试集进行了对比评测。我们将测试视频的第一帧截取出来作为静态图片，用测试视频作为驱动视频，生成一个与测试视频动作一样的视频，进而将测试视频与生成视频中的每一帧图像进行对比测评，最后取平均值。

下面我们逐一介绍这些测评指标的含义。

NIQE (Natural Image Quality Evaluator): NIQE 是一种用于评估自然图像质量的无参考指标。它基于“质量感知”特征，通过拟合到 MVG (Multivariate Gaussian) 模型中，从大量测试数据中提取的模型与当前视频的 MVG 模型进行距离计算，并将结果量化。由于特征是通过大量测试数据学习的，NIQE 不依赖于参考图像，因此是一种无参考评价方法。

PSNR (Peak Signal-to-Noise Ratio): PSNR 是峰值信噪比的缩写，用于评估图像的质量。该指标通过计算原视频和目标视频之间的均方误差 (MSE)，然后对结果取对数并乘以十来获得评判值。在 Python 中，已存在用于计算 PSNR 的函数，通常将图像转换为灰度图后直接传入该函数进行计算。

SSIM 与 MS-SSIM: SSIM (Structural Similarity Index) 是用于评估图像结构相似性的指标。它通过遍历像素点，从像素级别评估图像质量的好坏。MS-SSIM (Multi-Scale SSIM) 是在多尺度上计算结构相似性的指标，是 SSIM 的一种扩展方法，旨在更好地捕捉图像的细节信息。

LSE-C: LSE-C 通过唇形检测来衡量音频与图像同步的置信度。其值越高，表示同步效果越好。

LSE-D: LSE-D 通过唇形检测评估音频与图像的同步距离。其值越低，表示同步效果越好。

L1-loss: L1-loss 是通过计算像素点间的距离进行量化的指标。L1-loss 值越低，表示生成图像与原始图像的相似度越高，效果越好。

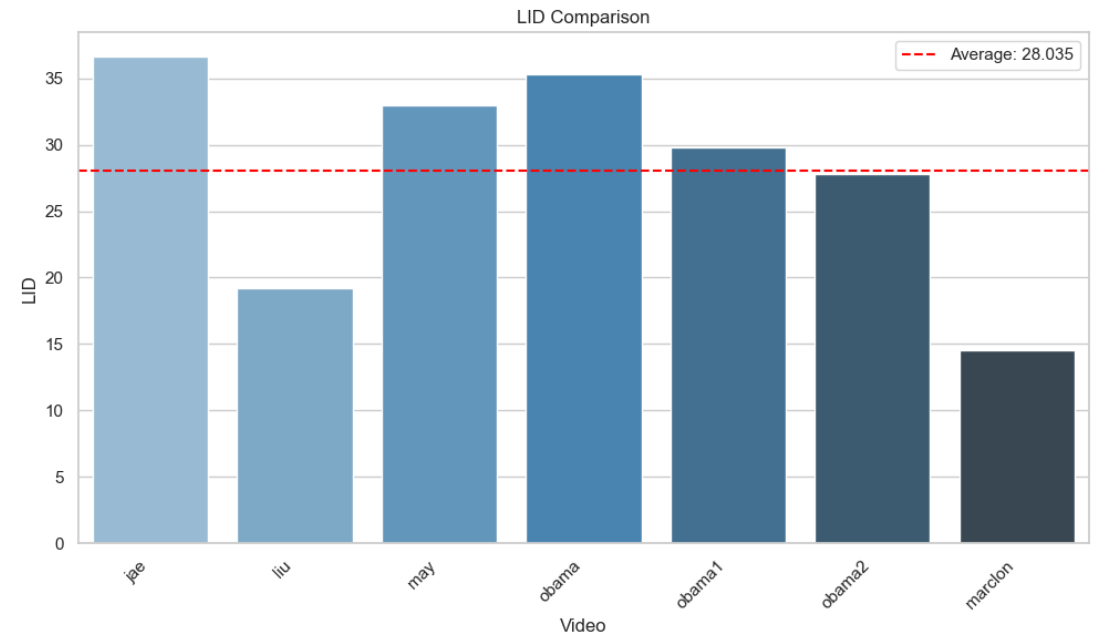
LPIPS (Learned Perceptual Image Patch Similarity): LPIPS 是基于神经网络的一种图像相似性度量方法，旨在符合人类视觉感知的标准。该方法通过从每个局部图像补丁中提取特征进行比对，而其损失函数使用的是感知损失函数，而非传统的均方差损失函数。

FID (Fréchet Inception Distance): FID 指标使用 InceptionV3 模型提取图像特征表示，并通过计算两个图像的均值和协方差的距离来定量评估图像的质量。FID 值越低，表示生成图像与真实图像的相似度越高，质量越好。

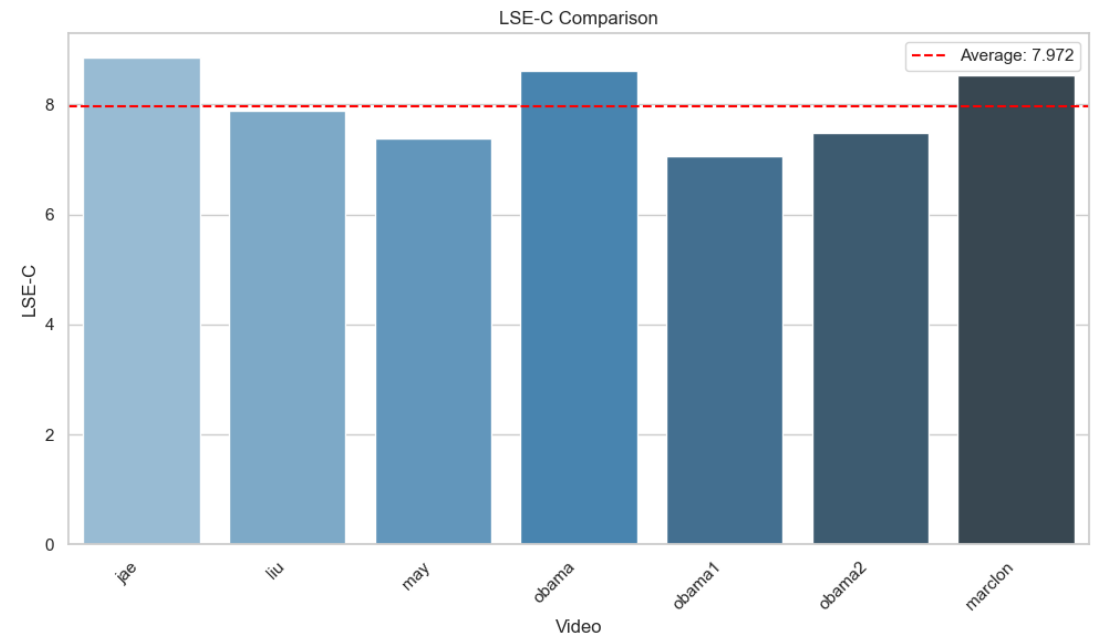
原视角生成评测结果如下：

评测指标	LID	LSE-C	LSE-D	PSNR	SSIM	L1-loss	MS-SSIM	NIQE	LPIPS
jae	36.615	8.851	4.231	24.213	0.854	9.751	0.904	50.583	0.065
liu	19.209	7.883	7.618	25.941	0.837	8.702	0.919	51.241	0.047
may	32.987	7.379	7.095	26.026	0.790	8.849	0.899	72.439	0.056
obama	35.336	8.613	6.331	19.332	0.619	14.739	0.669	31.247	0.102
obama1	29.791	7.054	7.867	25.734	0.862	8.002	0.931	34.714	0.038
obama2	27.793	7.494	6.392	24.898	0.869	8.655	0.924	41.490	0.040
marclon	14.517	8.532	4.675	26.367	0.884	6.954	0.934	25.050	0.039
平均数据	28.035	7.972	6.316	24.644	0.816	9.379	0.883	43.823	0.055

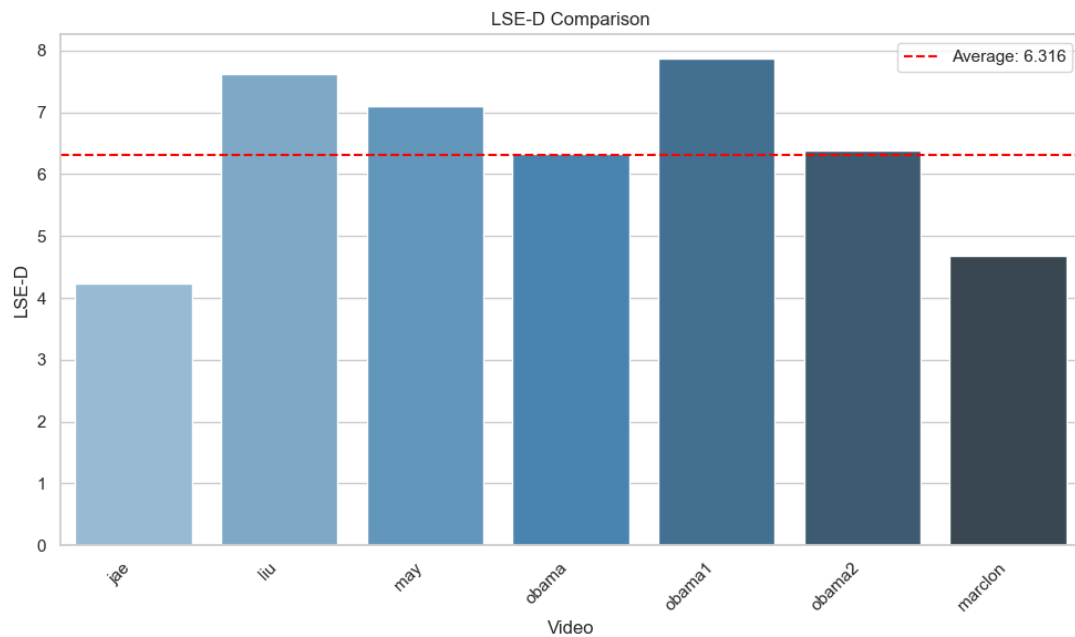
LID 的值在 28.035，根据这个指标的分档，其小于 50，是一个比较优秀的值，表示和生成图像和原图像表现较好，大多数的测试视频也在平均线上下浮动。



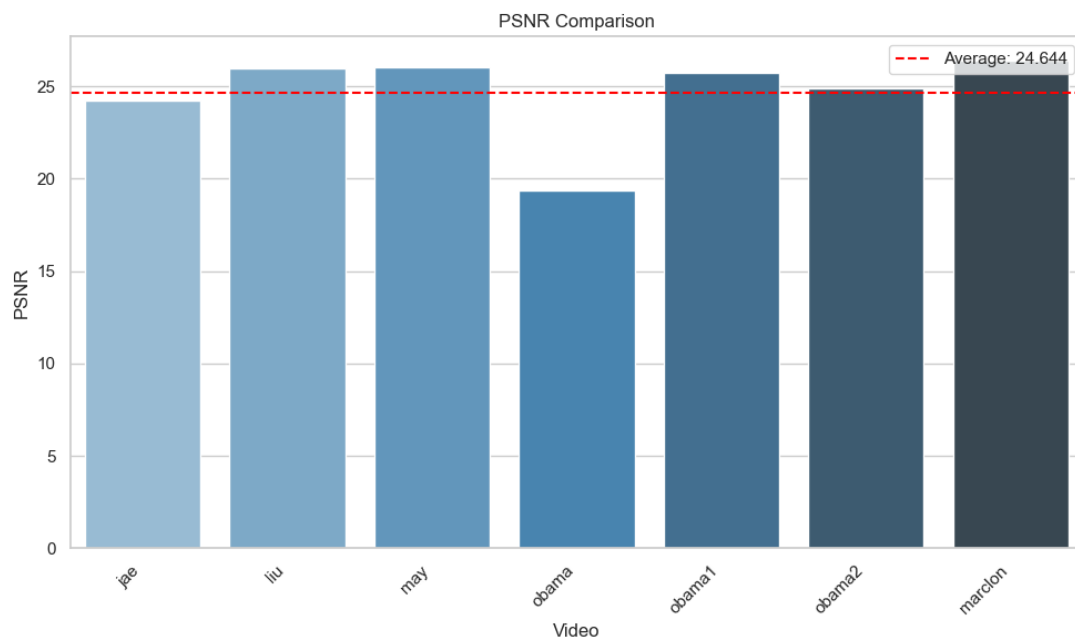
LSE-C 值为 7.972，且差距波动不大，证明其生成的视频在以唇形匹配为基础的音频同步的置信度比较稳定且比较出色。



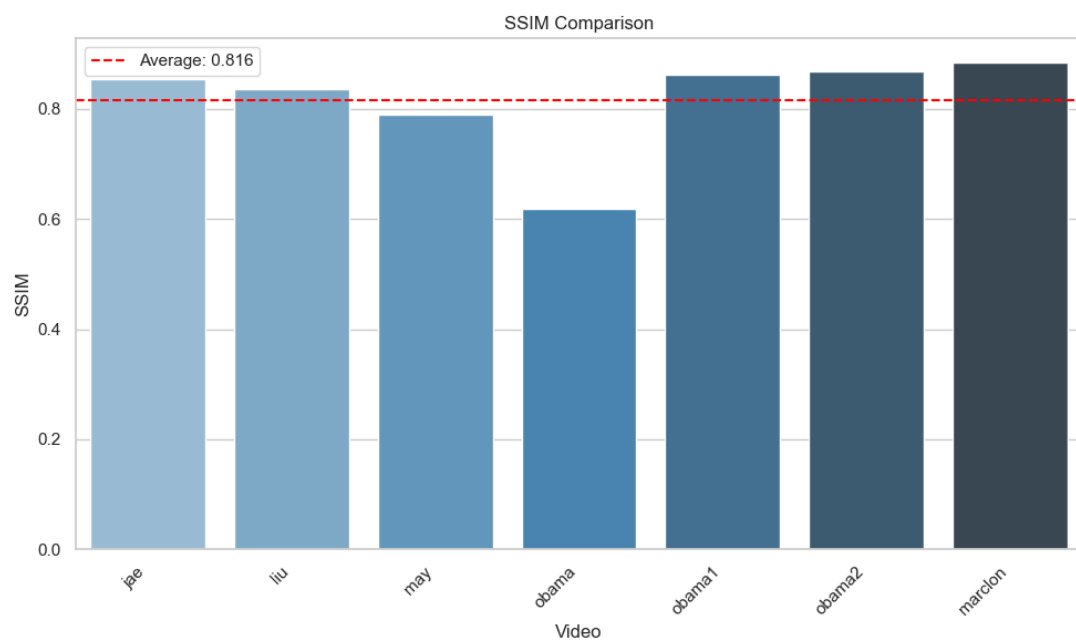
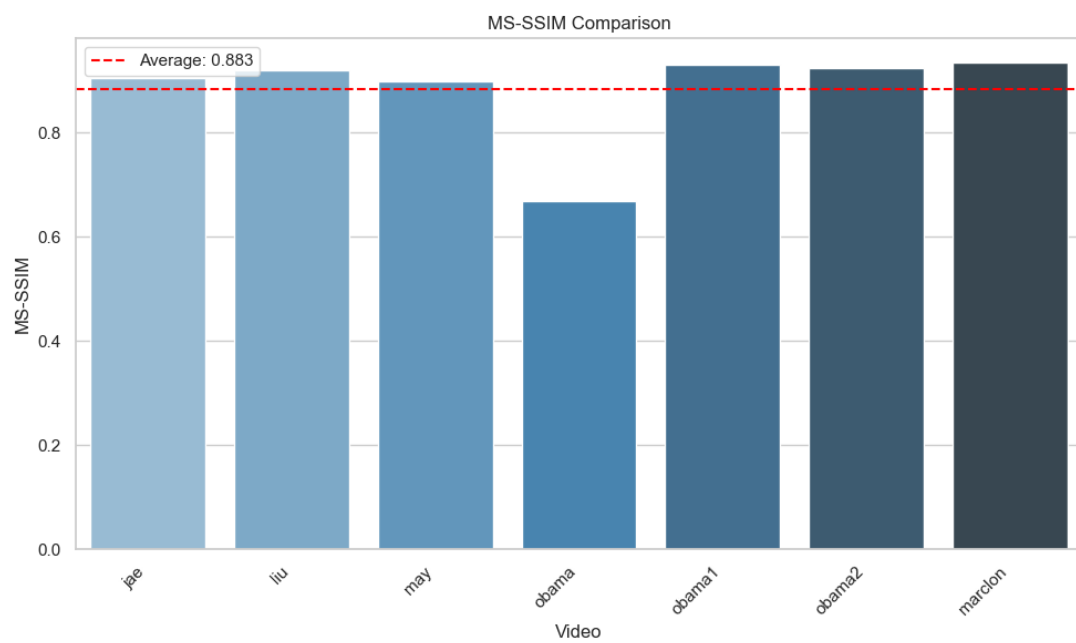
LSE-D 值为 6.316，且差距波动不大，证明其生成的视频在以唇形匹配为基础的音频同步的距离比较稳定，在比较好的范围。



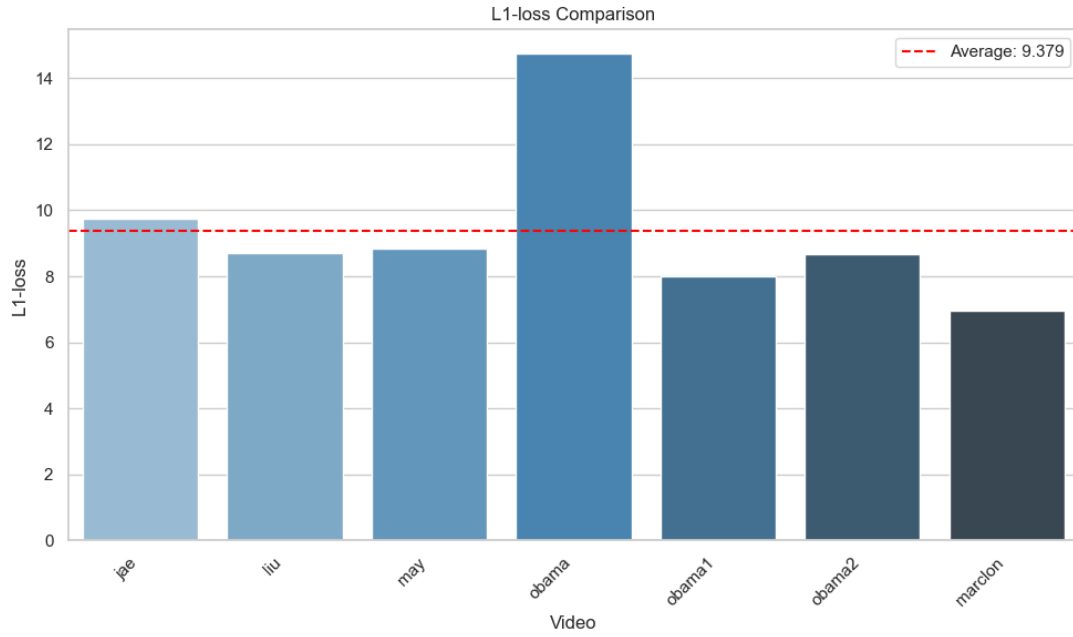
PSNR 为 24.644 dB，表示生成的视频还存在一定的噪声，但是我觉得这是因为测试集本身的差异导致的。因为我对测试集也跑了 PSNR，似乎这个指标的值也很高。



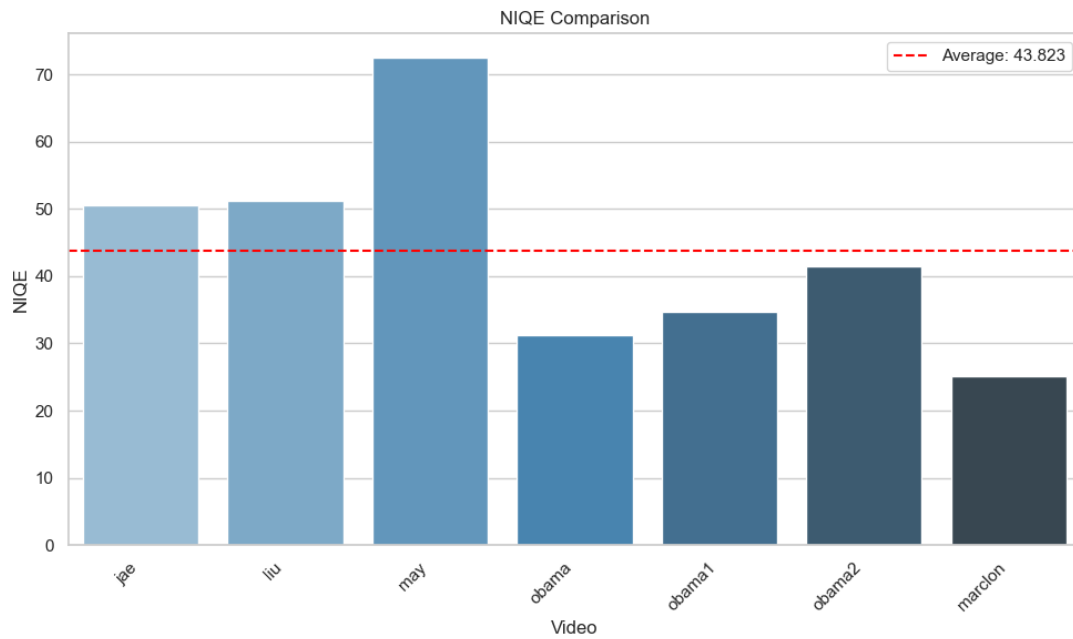
SSIM 与 MS-SSIM 分别为 0.816 与 0.883,这证明了两个视频在结构上的相似度比较高，保留了较好的细节信息。



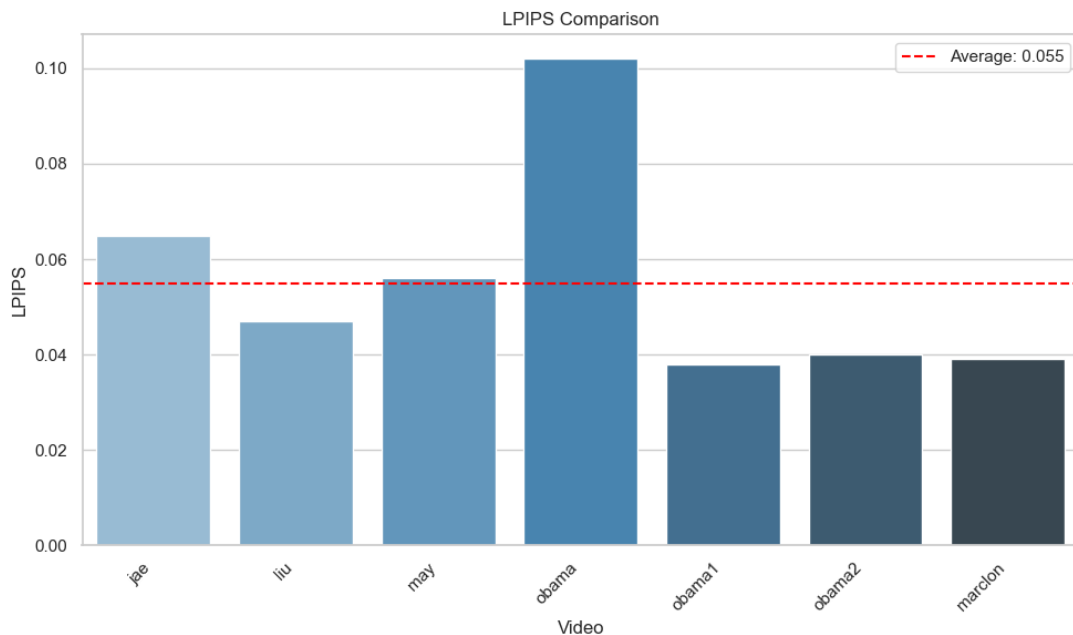
L1-loss 在 color 通道下的值为 9.379，从像素级别差异上来看，是一个比较好的表现。



NIQE 的平均评价结果为 43.823，这表明生成的视频和自然图像的差距属于中等水平，并且在不同的视频下起伏较大，有一个视频的 NIQE 高达 72.439，效果比较差，有一个又只有 25.050，效果比较好。可见和原先的图像质量有关联。



在 LPIPS 指标下，生成的视频有一个比较好的效果，仅有 0.055。这个指标下跑出来的值很小，证明生成的图像很适配人类的感知。



自由视角生成评估结果如下：

评测指标	LID	LSE-C	LSE-D	PSNR	SSIM	L1-loss	MS-SSIM	NIQE	LPIPS
jae	36.532	8.927	4.084	24.211	0.854	9.753	0.904	49.594	0.065
liu	19.847	9.014	6.495	25.942	0.837	8.702	0.919	52.339	0.047
may	33.479	7.799	6.539	26.026	0.790	8.852	0.898	71.980	0.056
obama	34.049	8.186	5.605	19.331	0.619	14.743	0.669	31.466	0.102
obama1	29.439	7.734	7.118	25.733	0.862	8.004	0.930	34.614	0.038
obama2	28.215	8.664	6.166	24.897	0.869	8.657	0.924	42.324	0.040
marclon	14.483	8.811	4.333	26.366	0.884	6.956	0.934	25.022	0.039
平均数据	28.006	8.448	5.763	24.644	0.816	9.381	0.883	43.906	0.055

自由视角选定的参数是(30, 10, 0)，得出的定量结果如图，将其与原视角进行对比：





















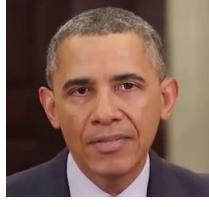








评测指标	LID	LSE-C	LSE-D	PSNR	SSIM	L1-loss	MS-SSIM	NIQE	LPIPS
自由视角	28.006	8.448	5.763	24.644	0.816	9.381	0.883	43.906	0.055
原视角	28.035	7.972	6.316	24.644	0.816	9.379	0.883	43.823	0.055




































可以看出很多指标其实变化并不大，但是我们发现有关于唇形匹配的指标似乎在自由视角下有更好的效果，可能是在偏转了角度之后其唇形展示出的面积更大导致的。其它指标的分析之前的大差不差。

2.3.2 定性评估结果

定性评估对理解模型性能十分重要，它能够展示模型生成视频与原始视频之间的差异。通过对比原始视频帧和生成视频帧的图片，我们可以直观地感受到模型在面部表情、细节恢复以及视觉一致性方面的表现。



生成的视频					
自由视角生成视频					
原始视频					
生成的视频					
自由视角生成视频					
原始视频					
生成的视频					

自由视角生成视频					
原始视频					
生成的视频					
自由视角生成视频					
原始视频					
生成的视频					
自由视角生成视频					

原始视频					
生成的视频					
自由视角生成视频					
原始视频					
生成的视频					
自由视角生成视频					

通过观察对比图，我们发现模型生成的视频在面部表情的恢复上表现优秀，尤其是在嘴巴和眼睛的动态表现方面，与原始视频几乎无异。然而，在快速运动的情况下，出现了轻微的模糊和失真。此外，模型在处理面部细节时，对于一些小范围的动作（例如微小的眼部眨动、嘴部变化等）可能会丢失或产生不自然的效果。

2.3.3 模型现存问题

首先，模型脚本本身可能存在问题。在 Docker 环境中，模型无法成功处理较大的视频文件，尽管设备上显示的 GPU 存在大量空闲空间。即使配置了 GPU 支持，仍然在处理较大视频时出现自动中断（killed）的现象。根据我的分析，这可能是由于设备资源不足或

资源分配问题所导致的。虽然 GPU 显存充足，但其他资源（如内存或存储）可能成为瓶颈。为了解决这一问题，可以尝试优化资源分配或将视频进行分段处理。

此外，生成的视频在某些指标下存在明显的抖动。在七个测试视频中，Obama 视频在各项指标上的表现与其他视频显著不同。通过初步分析，我推测这可能是由于该视频中的人物与背景的色调差异较低，导致模型在图像处理过程中，人物和背景的分辨率恢复不如其他视频。此问题尤为显著地体现在一些依赖像素精度的评测指标上，如 L1-loss 和 SSIM。这些指标对图像细节的变化非常敏感，低对比度的图像可能使得模型难以捕捉到细节，从而影响生成效果。

2.4 总结心得及对课程的建议（0.05*个人贡献度*项目分）

这个大作业中，最困难的部分实际上是环境配置。

首先是原项目本身，它几乎没有任何说明文档，README 中仅列出了两个 Python 包的版本信息，其中一个版本号还是错误的。为了搭建一个所有依赖包都兼容且能够成功运行原项目的环境，耗费了我们很大的精力。项目涉及许多依赖包，其中一些函数在新的版本中已经被移除或实现方式发生了变化，导致运行时频繁报错并互相冲突。小组的四名成员都尝试了环境配置，而我的过程尤为繁琐，由于中文临时文件的原因，我们还不得不修改脚本中的一些配置。最终，为了确保能在助教的电脑上成功复现，我们选择了组长电脑上的环境作为最终版本。

接下来是 Docker 环境的配置，这部分同样让人头疼。一开始，我们在电脑上安装了 Docker Desktop，但安装完成后电脑直接死机，导致 Docker Desktop 完全无法使用。随后，我们尝试在之前 WSL 的编译内核中重新配置了一套 Docker 环境，但 WSL 无法连接代理服务，导致无法下载 Docker 的镜像源。在查阅了大量资料后，我们找到了一些镜像加速站点，最终利用这些加速站点成功下载了镜像源，并封装好了所需的环境。

相比之下，编写评测脚本的过程要顺利许多。在参考一些资料和借助大模型的帮助下，我们一步步完成了调试，并成功获取了多项评价指标。这一过程不仅提升了我的技术能力，也让我们对项目有了更深入的理解。

参考文献：

- [1] WANG T C, MALLYA A, LIU M Y. One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing[J]. arXiv, 2021, arXiv:2011.15126. Available at: <https://arxiv.org/abs/2011.15126>.
- [2] NEWELL A, YANG K, DENG J. Stacked Hourglass Networks for Human Pose Estimation[J]. arXiv, 2016, arXiv:1603.06937. Available at: <https://arxiv.org/abs/1603.06937>.
- [3] DROBYSHEV N, CHELISHEV J, KHAKHULIN T, et al. MegaPortraits: One-shot Megapixel Neural Head Avatars[J]. Proceedings of the 30th ACM International Conference on Multimedia, 2022, 2663–2671. DOI: 10.1145/3503161.3547838. Available at: <https://doi.org/10.1145/3503161.3547838>.
- [4] SIAROHIN A, LATHUILIÈRE S, TULYAKOV S, et al. First Order Motion Model for Image Animation[J]. arXiv, 2020, arXiv:2003.00196. Available at: <https://arxiv.org/abs/2003.00196>.
- [5] THIES J, ZOLLHÖFER M, NIESSNER M. Neural Voice Puppetry[C]. SIGGRAPH, 2020. Available at: <https://arxiv.org/abs/1912.05566>.
- [6] JOHNSON J, ALRAHAMINEN A, FEI-FEI L. The Perceptual Losses for Real-Time Style Transfer and Super-Resolution[C]. ECCV, 2016. Available at:

<https://arxiv.org/abs/1603.08155>.

.....