武汉大学计算机学院 2022级中级项目实训周报

**专业班级：** 计卓班 **学号：2022302111279** **姓名：邱涵**

**项目名称： 缤果数智 小组成员：邱涵、陈震雄、陈晋宇、李佳隆**

**时间阶段：**第 2 周， 2024 年 7 月 8 日至 7 月 13 日

1. **本周计划和项目任务**

1、(算法组)继续进行Wav2Lip288x288的训练

2、(算法组)学习SadTalker原理

3、(算法组)训练SadTalker

4、(后端组)实现对于Edge-TTS等在内的第三方API调用

5、(前端组)学习Gradio

6、(前端组)使用Gradio实现对于TTS和数字人模型选择的界面设计

1. **本周进度/任务完成情况**

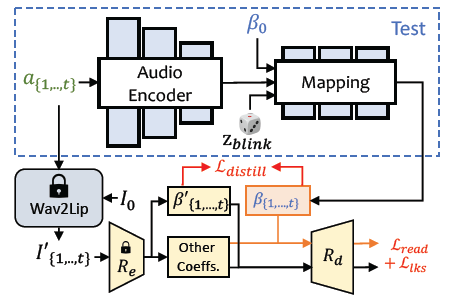
1、Wac2Lip288x288训练完成

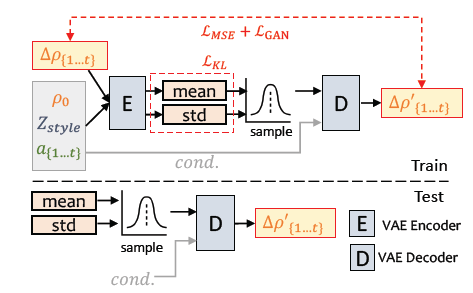
2、学习完SadTalker原理,如下：

SadTalker首先是生成3D面部。三维信息对于提高视频真实性至关重要，但之前的工作很少考虑三维空间。受单图像深度三维重建方法启发，论文这里使用预测的三维形变模型（3DMMs）作为中间表征，解耦三维脸部形状，并仅建模运动参数以实现身份无关的系数生成。

然后是通过音频生成运动稀疏。三维运动系数包括头部姿势和表情，学习所有系数会增加网络不确定性。SadTalker使用PoseVAE和ExpNet分别生成头部姿势和表情运动，以减少不确定性。

接着是比较重要的ExpNet，这个网络从音频中产生准确表情系数困难，因为音频到表情不是一对一映射任务。ExpNet通过第一帧表情系数联系特定人物，减少不确定性，并使用嘴唇运动系数作为目标。



然后是PoseVAE，PoseVAE是一个基于VAE的模型，学习真实、身份相关的风格化头部运动。它学习第一帧条件下的残差，生成更长、更稳定、更连续的头部运动，并增加音频特征和风格标识作为条件。

最后是 3D-aware面部渲染。生成三维运动系数后，通过三维图像动画器渲染最终视频。研究人员提出mappingNet学习3DMM运动系数和隐式无监督3D关键点关系，以提高运动自然度。训练包括自监督训练和在3DMM系数上微调mappingNet。

1. 训练SadTalker

主要是四个py文件来实现训练和推理的:

app\_sadtalker.py 实现了SadTalker系统的Gradio界面。它创建了一个用户友好的界面，允许用户上传源图像和音频，并提供各种设置选项如预处理类型和增强器。该脚本调用SadTalker的核心功能来生成动画视频，并在界面上显示结果。

inference.py 包含了SadTalker的主要推理逻辑。它处理输入的图像和音频，提取3DMM系数，将音频转换为动作系数，然后使用这些数据生成最终的动画视频。这个脚本是系统的核心，实现了从静态图像和音频到动态说话头像的转换过程。

launcher.py 负责设置SadTalker的运行环境。它检查并安装必要的依赖，确保系统运行所需的所有组件都已就绪。完成环境设置后，它启动SadTalker的Web界面，使用户可以通过浏览器访问和使用该系统。

predict.py 为SadTalker提供了一个预测接口。它定义了一个Predictor类，可以接受各种输入参数如源图像、驱动音频和增强器类型等。这个脚本运行SadTalker的核心功能并返回生成的视频，使得SadTalker可以作为一个独立的预测模块集成到其他系统中。

1. 学会了Gradio的基本的用法
2. 后端使用Python实现了对Edge-TTS、Paddle-TTS和ChatGPT在内等的多种大模型的API调用
3. 使用Gradio设计了TTS和数字人模型选择的基本界面
4. **下周的计划和项目任务**

1、将SadTalker、Wav2Lip等训练好的模型同前后端同学对接

2、优化模型训练，看是否能让模型训练效果更优

3、学习并训练音频克隆So-vits

4、学习并训练NeRFTalk

5、使用Gradio设计前端语音识别、LLM模型选择等的界面

6、使用Python实现前端调用后端和第三方接口，打通前后端

1. **存在的主要问题或特殊情况**

主要问题是组员之间交流对接代码不方便，接口API提供不符合规范等，下周这里需要仔细修改Bug。