深圳大学

SHENZHEN UNIVERSITY

计算机前沿技术

研究生创新示范课程研究报告

姓名： 邓文昌

学号：2310275031

2023 年 12 月 10 日

VideoReTalking: Audio-based Lip Synchronization for Talking Head Video Editing In the Wild

摘要

VideoReTalking使得说话者的面部表情、口型能与输入音频同步，同时即使在不同的情绪下也能产生高质量的假唱输出视频。VideoReTalking将这一目标分解为三个顺序任务：（1）具有规范表达的人脸视频生成；（2） 音频驱动的唇同步；（3）用于提高照片真实性的面部增强。给定一个会说话的头部视频，我们首先使用表情编辑网络根据相同的表情模板修改每帧的表情，从而生成具有规范表情的视频。然后该视频与给定的音频一起被馈送到唇同步网络以生成唇同步视频。最后，通过搭建身份识别人脸增强网络和后处理来提高合成人脸的照片真实性。VideoReTalking所有模块都可以在没有任何用户干预的情况下按顺序处理。此外，此系统是一种通用的方法，不需要针对特定的人重新训练网络。对两个广泛使用的数据集和野外实例的评估表明，在唇同步准确性和视觉质量方面，VideoReTalking的框架优于其他最先进的方法。

关键词：面部动画、视频合成、音频驱动生成

1 引言

这篇文章提出了一种新的系统来编辑会说话的嘴唇，以使输入音频与更稳定的嘴唇同步结果和更好的视觉质量相匹配。以前的作品将视频中的原始帧视为头部姿势参考。然而，我们发现嘴唇生成对这些参考非常敏感，并且直接使用原始帧作为嘴唇生成的基础往往会产生不同步的结果。为此，如图 1所示，我们采用了一种分而治之的策略，首先消除面部表情，然后使用修改后的帧作为嘴唇生成的姿势参考，考虑到所有参考人脸现在都有相同的规范表情，这更准确。最后，与之前经常产生低分辨率和模糊结果的工作相比，我们通过所提出的身份感知增强网络和修复来产生照片逼真的结果。

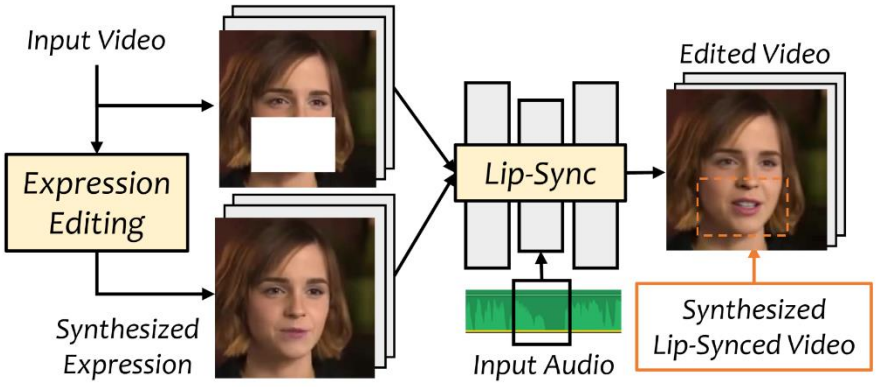


图 1 通过表情编辑和唇同步网络，通过输入音频修改原始视频并生成唇同步视频

具体而言，在给定任意通话视频的情况下，我们首先裁剪人脸区域，并通过深度神经网络提取3D变形模型（3DMM）的姿势和表情系数（Deng et al.，2019b[1]）。然后，我们使用具有标准中性模板表达的3DMM的参数，并通过类似于（Ren et al.，2021[2]）的语义引导表达重演网络重新生成视频。通过这样做，我们获得了一个在所有帧中具有相同规范表达的视频，它们将被视为我们唇同步网络的结构参考。有趣的是，我们还可以通过改变表情

模板来合成不同情绪的会说话的头部视频。例如，通过改变表情模板的嘴唇形状以匹配“快乐”情绪，这种嘴唇形状将在嘴唇同步网络中被考虑在内，导致谈话头部视频呈现出相同的情绪。

在表情中和之后，使用合成的表情作为条件结构信息，应用唇同步网络来合成逼真的下半人脸。具体而言，我们设计了一个以快速傅立叶卷积块（Chi et al.，2020[3]）为基本学习单元的沙漏型网络。关于音频注入，我们使用自适应实例归一化（AdaIN）块（Huang和Belongie，2017[4]）来调制全局中的音频特征。最后我们使用预先训练的唇同步鉴别器来确保视听同步性。

尽管之前的步骤可以合成嘴唇形状相对准确的会说话的头部视频，但视觉质量仍然受到低分辨率训练数据集的限制。为了解决这个问题，我们设计了一个保留身份的人脸增强网络，通过渐进训练产生高质量的输出。在通过人脸恢复方法（Yang et al.，2021[5]）增强的LRS2数据集（Afouras et al.，2018[6]）上训练增强网络。我们还应用StyleGAN先验引导的面部修复网络（Wang et al.，2021c[7]）来去除牙齿周围的视觉伪影。

在几个现有的基准以及野外视频中进行了广泛的实验来评估我们的框架。结果表明，所提出的系统可以产生比以前的方法高得多的视觉质量的视频，同时提供准确的嘴唇同步。

2 相关工作

根据输入的语音音频编辑会说话的头部视频的任务具有重要的现实应用，例如将整个视频翻译成不同的语言，或者在视频录制后修改语音。这项任务被称为视觉配音，在之前的几项工作中已经进行了研究（Suwajanakorn等人，2017[8]；Wen等人，2020[9]；Thies等人，2020[10]年；Prajwal等人，2021[11]），通过修改面部动画和情绪以匹配目标音频来编辑输入的会说话的头部视频，同时保持所有其他动作不变（如图2所示）。一些方法可以在特定的说话者身上获得令人满意的结果，但需要在目标说话者的谈话语料库上进行训练以获得个性化模型，而这并不总是可用的。另一方面，目前的通用方法会产生模糊的下脸（Prajwal et al.，202011）或不准确的嘴唇同步（Song et al.，2022[12]），这在视觉上是侵入性的。这些方法也不支持情绪编辑，这在改变语音内容时通常是可取的。

2.1 随机受试者方法

随机受试者方法旨在构建一个不需要针对不同身份进行再培训的通用模型。通过修复重建下半脸最近很流行。例如，LipGAN（KR等人，2019[13]）设计了一个神经网络来填充下半张脸作为姿势先验。Wav2Lip（Prajwal等人，202011）使用预先训练的SyncNet作为嘴唇同步鉴别器来扩展LipGAN（Chung和Zisserman，2016[14]），以生成准确的嘴唇同步。基于Wav2Lip，SyncTalkFace（Park et al.，2022[15]）涉及音频嘴唇记忆，以隐式存储嘴唇运动特征，并在推理时检索它们。另一类方法首先预测中间表示，然后通过图像到图像的翻译网络合成照片逼真度结果，例如，面部标志（Xie et al.，2021[16]）和基于3D面部重建的面部标志（Song et al.，2022[17]）。然而，所有这些方法都在努力合成具有可编辑情感的高质量结果。

2.2 特定人方法

性化的视觉配音比通用的更容易，因为这些方法仅限于已知环境中的特定人。最近的视觉配音方法侧重于从音频中生成中间表示，然后通过图像到图像的翻译网络渲染照片逼真的结果。例如，一些作品（Thies et al.，2020[18]）专注于来自音频特征的表达系数，并通过图像生成网络渲染照片逼真的结果。通过投影3D渲染的人脸，face landmarks（Lu et al.，2021[19]）和边缘（Ji等人，2021[20]）也是流行的选择，因为它包含更稀疏的信息。此外，基于3D网格的方法（Lahiri et al.，2021[21]）和基于NeRF（Mildenhall et al.，2020[22]

）的方法也很强大。尽管这些方法可以合成照片逼真的结果，但它们的应用相对有限，因为它们需要在特定的人和环境上重新训练模型。

2. 3 基于音频的单图像人脸动画

与视觉配音不同，单图像人脸动画旨在通过单音频驱动生成动画，同时也受到了视频驱动人脸动画的影响。例如，（Song et al.，2018[23]）使用递归神经网络从音频中生成运动，（Zhou et al.，2019[24]）通过对抗性表示学习将输入分解为主题相关信息和语音相关信息。（Vougioukas等人，2020[25]）将音频视为潜在代码，并通过图像生成器驱动人脸动画。在这个任务中，中间表示也是一个流行的选择。ATVG（Chen et al.，2019[26]）和MakeItTalk（Zhou et al.，2020[27]）首先从音频中生成面部标志，然后使用标志到视频网络来渲染视频。（Zhang et al.，2021a[28]）从音频中预测3DMM系数，然后将这些参数传递到基于流的扭曲网络中。（Wang et al.，2021a，b[29]）借鉴了视频驱动人脸动画的思想（Siarohin et al.，2019[30]）。

3 本文方法

3.1 本文方法概述

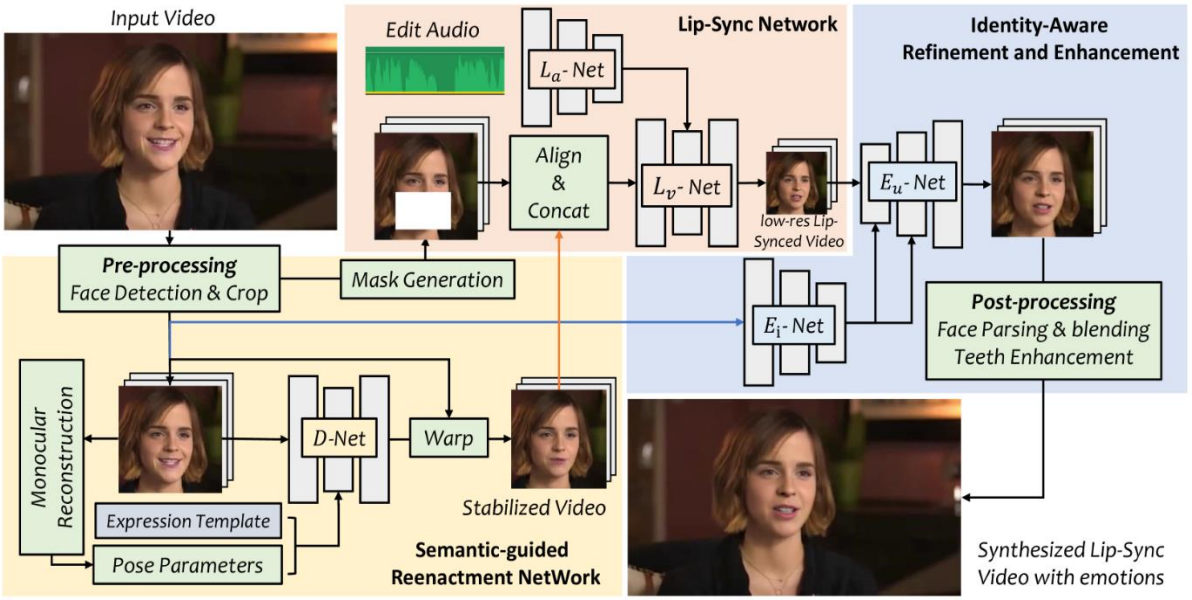


图 2 Frame Work

本文主要搭建了三个网络，它们的训练是相互独立的，对于D-net，它的目标是编辑人的表情，使得视频中人的表情变得“稳定”，不会出现过大的变化。对于L-net, 它的作用是根据输入人讲话的音频生成的人对上音频口型视频。最后了使得生成视频更加清晰，训练E-net生成高质量的视频。

如果仅使用L-Net，则存在两个主要问题。第一种是由原视频参考引起的信息泄漏，其中生成的唇缘仍然严重依赖于参考。另一个是可能会产生低视觉质量，因为当前的大型会说话的头部数据集的分辨率较低。

3.2 D-Net

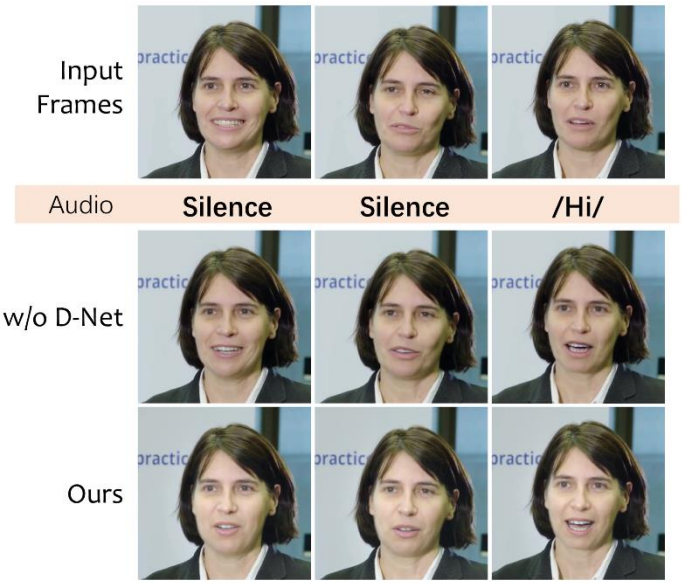


图 3 information leak

如果直接使用原视频作为L-net的参考，会导致什么问题？我的个人理解是虽然输入网络的video嘴巴部分被遮掩，但是视频中的谈话者面部透露出奇怪、极端的表情会可能会泄露原视频说话内容信息，导致网络以此作为参考从而导致生成的视频和音频衔接得不自然。这现象被文中称为information Leak。图 3形象地解决了这个问题。为了解决这个问题作者不将原视频作为参考，而是搭建了D-net冻结人的表情，将此处理后的视频传入L-net作为参考。根据我的理解作者在D-net模块完成了一下步骤。

(1):人脸检测+裁剪,然后对检测人脸Landmark， 对landMark进行Savitzky－Golay滤波器对其进行平滑处理,其中使用眼睛中心和鼻子的关键点作为面部对齐的锚点。

(2):单目人脸重建：从第一步得到的结果每个帧进行单目人脸重建从而获得提取姿势和表情系数。若想要修改表情，提供表情模板，若不提供模板这表情薄板置为0。

(3):第一步和第二步结果传入D-net相应分支，生成稳定的表情。

文章并没有没详细说明D-net的结构，经过查阅代码和参考文献D-net的结构类似于PIRenderer(Ren et al.,2021[31])的网络结构，图 4展示了它大致的输入输出结构。即D-net包含着三个子网络，分别是Mapping Network，warping network，editing network，每个网络是编码器和解码器的结构。与PIRenderer不同的是D-net得mapping net只需要传入Expression模板，人脸重建系数和epression传入mapping network后转变为隐藏变量Z,隐藏变量Z会被注入warping network，editing network的ADAIN层当中。warping network产生光流，根据光流修改人脸位置、表情。Editing network使得warping network产生的视频更加清晰。

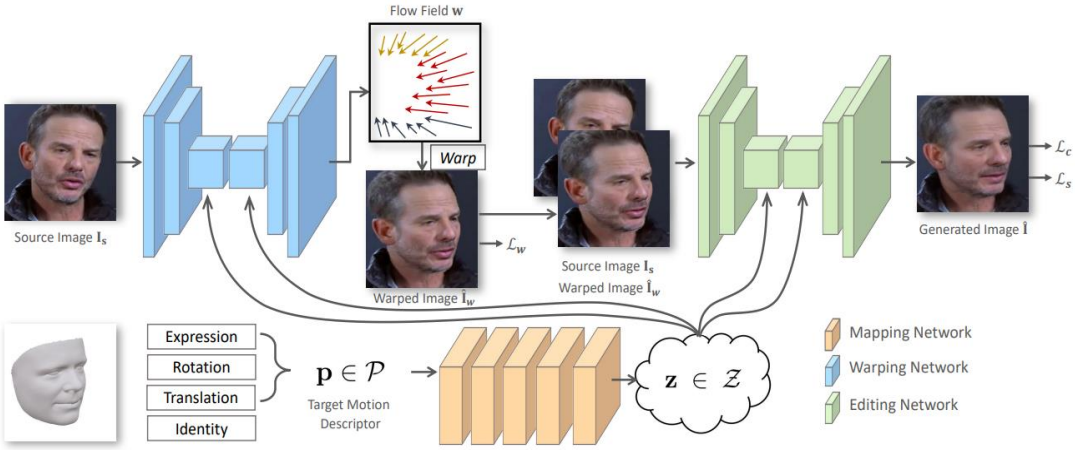


图 4 PIRender 网络结构

对于warping network，editing network使用perceptual loss L_{D_W}来监督学习，用来期望产生视频与原视频在感知上是同一个对象。L_{Dw}被定义为

$$LDW = L_{PERCEPTUAL} = \sum_l ||f_{vgg}^l(I_{gt}) - f_{vgg}^l(I_{D_W})||_2, \tag{1}$$

其中f_{vgg}^l是VGG-16网络[32]的第l层，I_{gt}, I_{D_W}分别为GroundTrue原来的图像和warping network生成的图像。而对于Editing network使用内容损失L_c和风格损失L_s监督它的训练。这两者被定义为

$$\begin{aligned} L_c &= \sum_l ||f_{vgg}^l(I_{gt}) - f_{vgg}^l(I_D)||_2, \\ L_s &= \sum_l ||G(f_{vgg}^l(I_{gt})) - G(f_{vgg}^l(I_D))||_2, \end{aligned} \tag{2}$$

其中I_D为editing network生成的图像，G是计算的常用于风格迁移的Gram矩阵。使用权衡系数调整两者比重，那么editing network损失函数为

$$L_{D_e} = \lambda_s L_s + \lambda_c L_c \tag{3}$$

3.3 L-net

图 5中expression Editing就是训练完成的D-net在工作，所以在训练L-net得完成D-net的训练（见3. 2）。

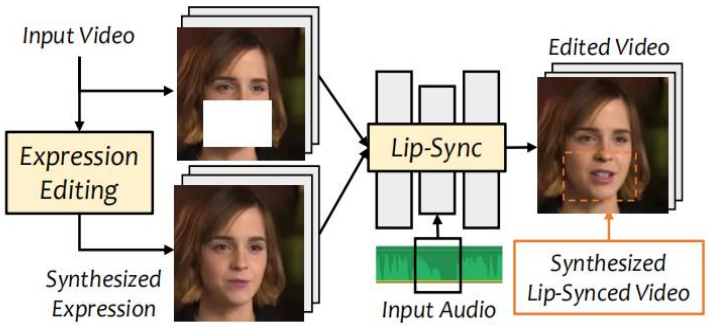


图 5 从D-net到L-net

如图 6所示，L-Net包含两个子网络，La和Lv，分别用于音频和视频处理。我在复现细节中展现L-net的具体结构。数据传入网络之前做了以下处理:

将音频转换为梅尔频谱：截取窗口大小0. 2s的音频（256维），对次拼音进行梅尔频谱转换，提取的特征尺寸大小为 80×16。

原视频检测出人脸并将下半张脸裁剪掉的masked的视频，从D-net获得处理过表情的视频作为参考。 图 6中最下面的人脸即为D-Net编辑过的人脸。最后将遮掩的视频和D-net处理的视频进行对齐拼接。

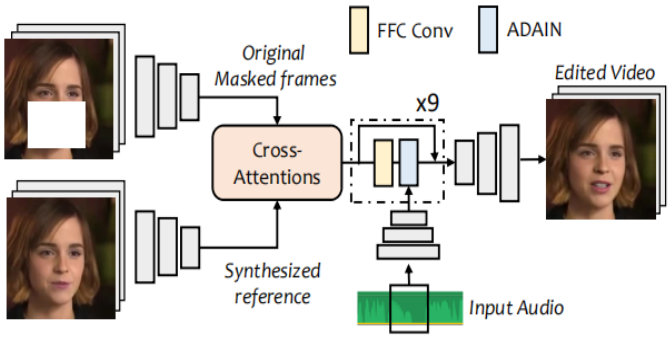


图 6 L-net 的大致网络结构

Q:如何控制生成的视频中人和原视频的姿势、状态等内容一致？

$$L_1 = ||I_{gt} - I_{OR}||_1 \tag{4}$$

$$L_{PERCEPTUAL} = \sum_l ||f_{vgg}^l(I_{gt}) - f_{vgg}^l(I_{OR})||_2, \tag{5}$$

其中I_{OR}是L-net生成的视频，文章使用了L1范数在RGB空间中控制内容损失和同时也使用了D-net中的一样的感知损失$L_{PERCEPTUAL}$。

Q:如何控制生成的视频中的嘴型能对得上输入音频？

$$L_{sync} = -\frac{1}{N}\sum_1^N \log(P_{sync}),$$
$$P_{sync} = \frac{v \cdot a}{\max||v||_2 \cdot ||a||_2}, \tag{6}$$

其中P_{sync}直接从训练好的SyncNet[33]中获得，输入视频帧和帧间隔内音频信号经过SyncNet分别解码得到v，a，然后公式(6)就P_{sync}反应视频中的口型和音频是对得上的概率。最后L-Net的loss函数表达如下

$$L_L = \lambda_1 L_1 + \lambda p L_{PERCEPTURAL} + \lambda_{SYNC} L_{SYNC} \tag{7}$$

3.3 E-Net

L-Net的生成的结果仍然不完美，原因有两个：很难在高分辨率的会说话的头部数据集上训练模型且还没有公开的大规模高分辨率会说话的头部数据集。所以L-Net使用都是低像素的训练集。为了获得高分辨率的会说话的头部数据集和用于上采样的对齐域，文章首先使用基于GAN先验的人脸恢复网络来增强低分辨率数据集（Yang et al.，2021[34]）。若直接使用这个人脸恢复网络作用在L-net上会导致牙齿和面部模糊也可能导致生成身份特征发生变化[35]。所以作者使用了以下流程：

（1）GAN人脸恢复网络增强原始数据集，使得单个视频宽高从96×96变为384×384。

（2）增强的数据传入Ei-Net得到identity feature code。

（3）增强的数据传入经过下采样重新变为96×96传入已经训练好的L-Net产生低像素视频。

（4）第（2）步得到的identity feature植入Eu-Net中间层以及第（3）步产生的低像素视频输出Eu-Net生成对高像素的嘴型视频。

E-net的损失函数分为三部分。第一部分的loss负责控制模型生成的图像能够内容与原视频中的一致，这部分loss的意图与训练L-Net用到的内容损失差不多，换汤不换药，也是L_1和$L_{PERCEPTUAL}$来控制，分别表达如下

$$L_1 = ||I_{gt} - I_{HR}||_1 \tag{8}$$

$$L_{PERCEPTUAL} = \sum_l ||f_{vgg}^l(I_{gt}) - f_{vgg}^l(I_{HR})||_2, \tag{9}$$

其中，I_{HR}是模型生成的高像素图像。为了控制生成人身份特征不损失，作者使用了预训练的人脸识别网络ArcFace(Deng et al., 2019a[36])，通过输入生成图像和原始图像来借助此网络判别是不是同一个人。

$$L_{id} = |||f_{arcface}(I_{gt}) - f_{arcface}(O_{HR})||_2 \tag{10}$$

其中$f_{arcface}$是人脸识别网络的特征编码。最后为了控制生成内容的真实性。作者还定义了一个判别器D（对每一个样本生成一个概率，概率越大任务），E-net通过与判别器的对抗提高生成样本的真实性，对于生成器E-Net和判别器损失D被定义为

$$L_{adv}(G_E, D) = E_{I_{gt}}[\log D(O_{HR})] + E_{O_{HR}}[\log(1 - D(G_E(O_{HR})))] \tag{11}$$

其中G_E就是E-net生成器。最后对E-net网络优化目标表达为

$$(G_E^*, D) = arg \min_{G_E} \max_D \lambda_1 L_1 + \lambda_p L_{PERCEPTUAL} + \lambda_{adv} L_{adv} + \lambda_{id} L_{id} \tag{12}$$

图 7展示了E-net与其他方法的结果对比。可以看到，GFPAGAN身份特征被丢失了很多，而GPEN能够生成高像素的图像但是人物的牙齿相对来说模糊了一些。

图 7 E-Net与其他方法对比

4 复现细节

4.1 与已有开源代码对比

1．源码没有有提供训练模型的代码，我根据文中给出的Loss函数，编写trainer，以尽可能实现D-net模型训练。

2. 文章使用了第三方的预训练好的模型来指导模型训练或者做数据预处理与增强，我根据文章的参考文献尽最大可能把代码的依赖的算法和模型提供出来。例如使用预训练的SyncNet指导L-Net的学习。

3. 根据文章附录给出的具体模型结构，我检查源码并添加模型以及检查中代码中有bug的地方，针对结果中出现的问题提供相应的修复模型。

源码地址：[OpenTalker/video-retalking: [SIGGRAPH Asia 2022] VideoReTalking: Audio-based Lip Synchronization for Talking Head Video Editing In the Wild (github.com)]

4.2 实验环境搭建

1. Anaconda安装创建了创建以下虚拟环境：python3. 8. 14并安装库：

ffmpeg torch==1. 9. 0+cu111torchvision==0. 10. 0+cu111、basicsr==1. 4. 2

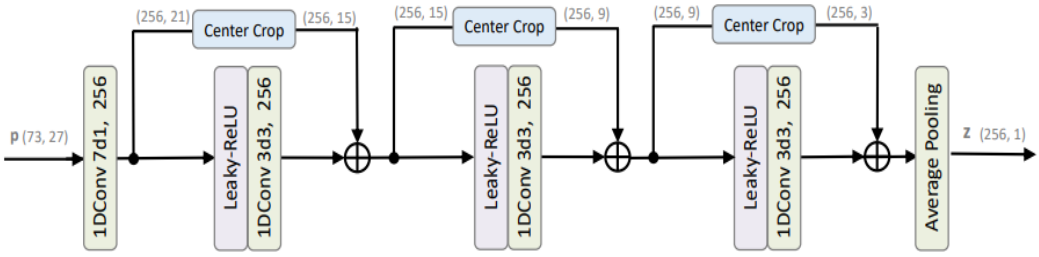
kornia==0. 5. 1 face-alignment==1. 3. 4 ninja==1. 10. 2. 3 einops==0. 4. 1

facexlib==0. 2. 5 librosa==0. 9. 2 dlib==19. 24. 0 gradio>=3. 7. 0 numpy==1. 23. 4

2. 平台window10，显卡：GTX1060×1，cpu:intel i9

4.3 模型结构

D-Net模型结构：类似于pirender网络的mapping netwok, 其次是warng network，最后是editing network。它们具体结构



(a) The Mapping Network

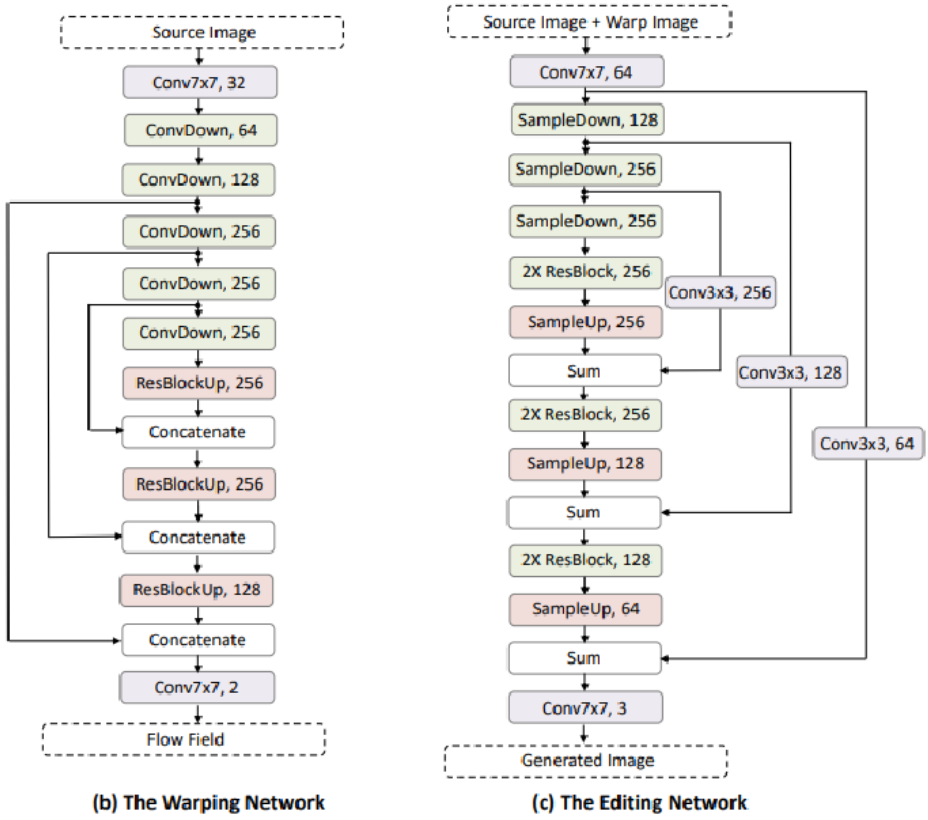


图 8 PIRender网络结构(D-Net参考结构)

使用vox celeb数据集进行训练。其中 the mapping network和the warping network训练了200k iterations，在每个卷积层之后应用AdaIN块来注入mapping network输出的z。然后训练整个网络200k iterations。优化器使用Adam，学习率初始化为e-4，对于整个网络的损失函数如公式(7)被定义为两部分。其中λc和λs分别为1和250。

L-Net模型结构：L-Net是多分支结构，其中一个分支处理音频的L_a，另外两个分支处理视频分，别是被遮掩口部的视频和参考视频（被D-net处理表情后的视频）。L-Net使用了两各相邻cross attention:将masked视频经过encoder得到的特征，作为Q, K, 而参考对应的特征充当V。同时使用最新的FFC快速傅里叶卷积FFC来提高网络获取全局信息的能力[37]。图 9展示了L-net的结构图, 而图 10进一步讲述每个块的详细结构。

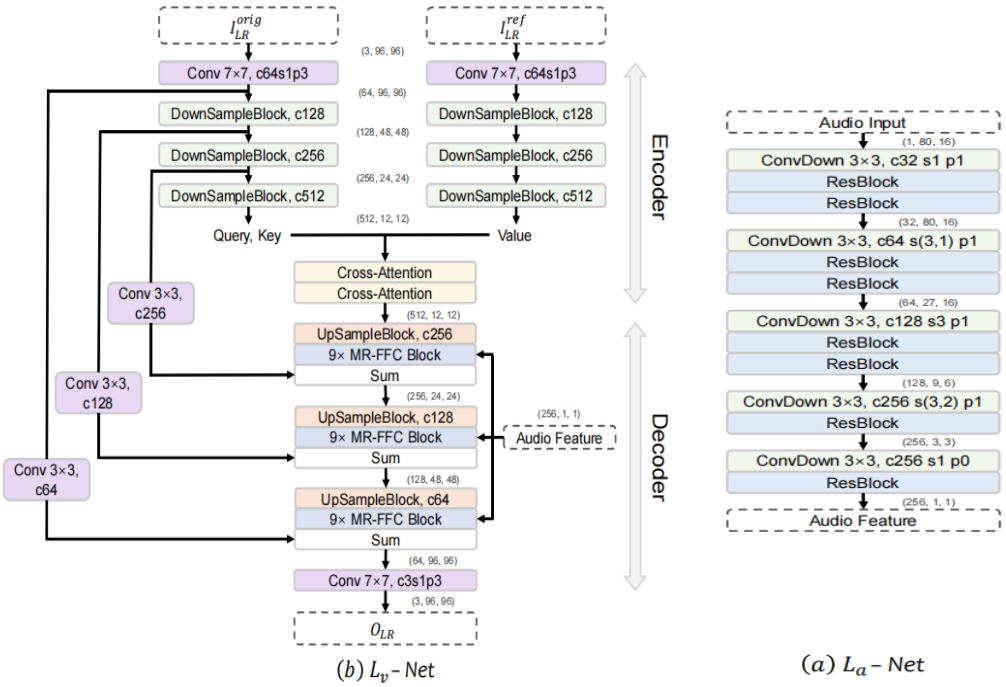


图 9 L-net网络结构

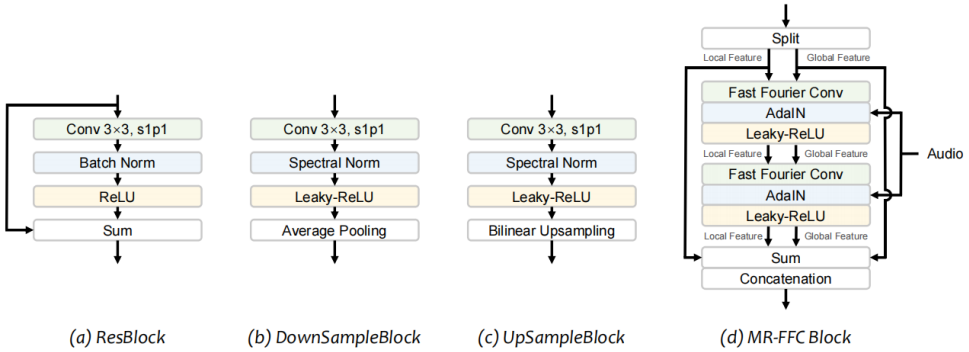
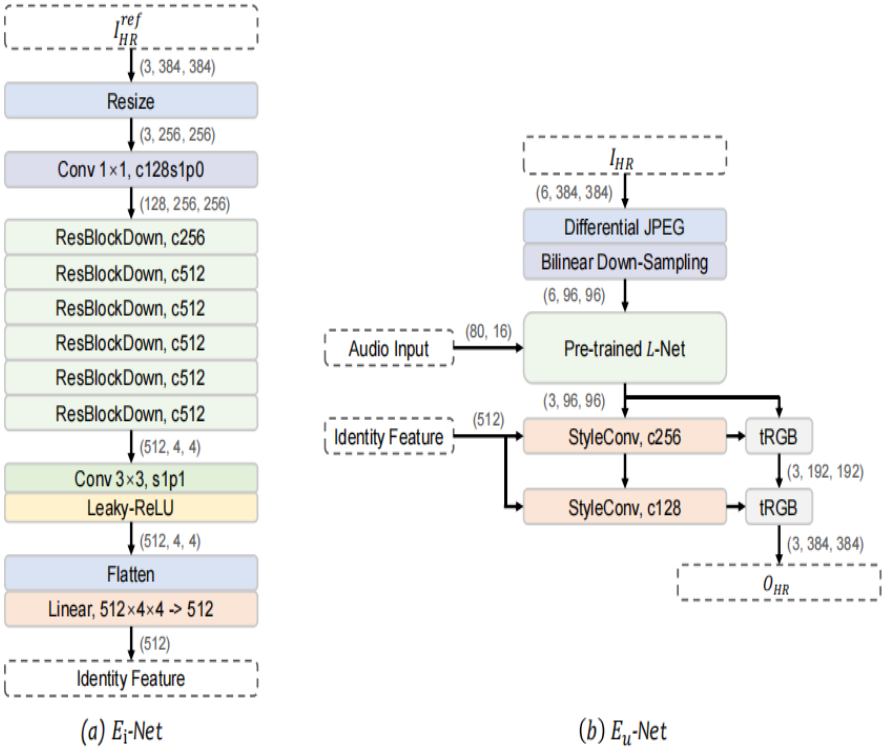


图 10 L-Net 网络中每个块使用的具体结构

在训练阶段，输入大小为$R^{5\times6\times96\times96}$，5表示输入帧由五个连续帧组成，其中五个帧是从同一视频中随机选择的参考；6是masked的图像和参考图像在RGB通道进行拼接而来；96是视频图像的高宽。使用Adam优化器在LRS2数据集上进行400k次迭代来训练，学习率为1e−4。公式(7)为L−net的loss函数，λ_1，λp，λ_{SYNC}分贝被设置为1, 1, 0. 3。

E−net网络结构包含一个身份编码器Ei−Net和一个超分辨率模块Eu−Net。在Eu−Net中，与L−Net类似，我们将视频帧的连续五帧（屏蔽的图像和D−net编辑图像）和从同一视频中随机选择的引用馈送到L−Net。然后，这些图像进行下采样，并发送到预训练的L−Net进行唇同步。超分辨率模块EU−Net使用类似的块对低分辨率结果进行上采样。构建StyleConv块和tRGB块以学习高分辨率结果。我们首先调整从中随机选择的高分辨率参考帧的大小。然后，下采样层将用于提取的高级特征到一个维度向量。图 11和图 12为D−net的具体网络结构。



(a) E_i-Net　　(b) E_u-Net

图 11 E-net网络结构

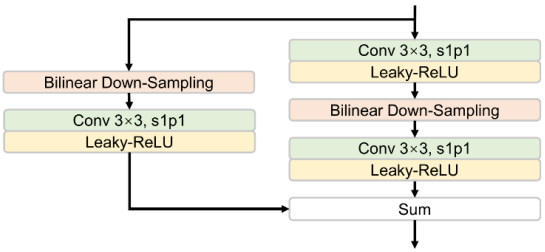


图 12 E-net中resblock结构

公式(12)是E-Net的Loss function，其中λ_1，λ_p，λ_{adv}，λ_{id}分别设置为0.1，2，100，0.4。

4.4 创新点

原文中D-net模型参数数量庞大，无论是训练还是推理阶段都需要大量的时间，直接取视频第一帧或者浅层三维人脸重建代替可以加速模型的训练和推理。

若输入的音频含噪声，即使音频没有谈话内容，那么生成的视频中人的嘴型依然处于说话状态，在模型前加入denoise模型可以缓解此问题。

5 实验结果分析

5.1D-net的训练

数据处理：如图 13 所示，首先检测面部标志，我使用的shape_predictor_68_face_landmarks模型。利用时间Savitzky－Golay滤波，然后使用眼睛中心（绿色）的关键点以及鼻子（绿色部分）作为用于面部对准的锚点，红色部分是受用Savitzky－Golay滤波后眼睛关键点的检测。

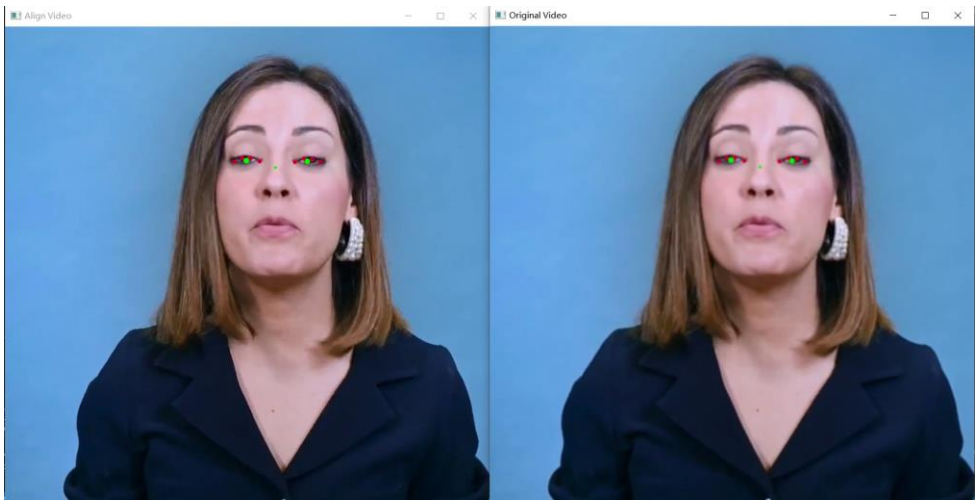


图 13 人脸关键点检测

使用人脸检测然后将人脸裁剪下来，使用matplotlib的可视化如图 14

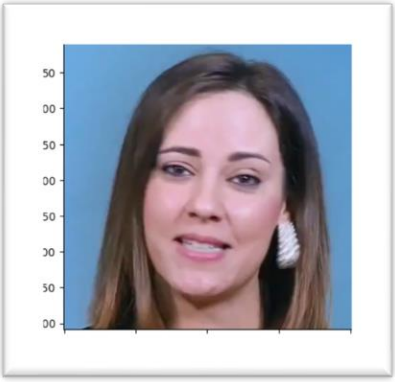


图 14 人脸检测+裁剪

然后经过3dmm重建得到视频每一帧的系数，用smile模版替换每一帧系数矩阵前64行（只修改下半张脸）。最后定义内容损失函数和风格损失，与原文不同的是内容损失被定义为输出图像和原图像在vgg16最后一个非全连接层的特征输出的L2距离，风格损失被定义输出图像和原图像在vgg16最后一个非全连接层输出的Gram矩阵的L2距离。

优化器与原文描述一致，使用Adam和学习率1e−4。训练集：人脸口罩训练集（爬虫获取），迭代300次，Batchsize=10。

5.1 D-net的inference

将此结果和视频第一帧作为source送入D-net（已训练）。图 15为 D-net在此设置下输出的图片。

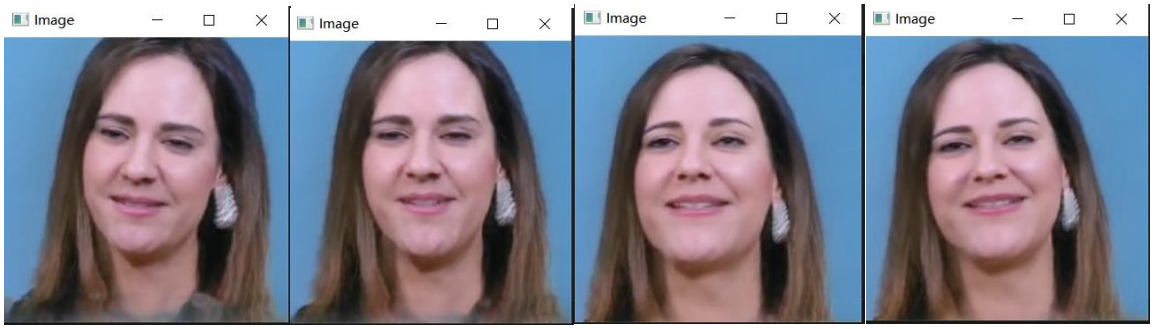


图 15 第一帧作为D-net的source输出的结果（1）第1帧 （2）第50帧 （3）第100帧 （4）第150帧

不采用第一帧作为source而是采用视频每一帧作为每次输出的source结果如图 16所示。

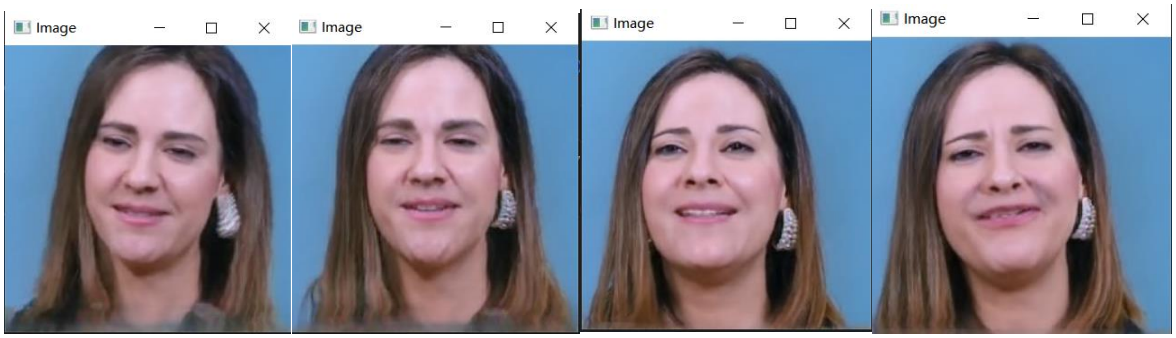


图 16 每一帧作为D-net的source输出结果（1）第1帧 （2）第50帧 （3）第100帧 （4）第150帧

通过对比两次结果结果，可以发现除了第一帧完全相同以外，第二种设置下模型输出视频中女子的嘴轻微地在动且人脸更加清晰，原因很明显，原视频女子正在谈话，所以每一帧女子的嘴唇都会不同，而D-net更想要表情稳定的结果，不期望嘴部出现太大变化，同时实验结果解释了为什么文章倾向于使用第一中设置。

5.2 发现

当输入的纯噪声音频且无人声时，人的嘴型依然在动，原因就在训练L-net使用的loss function仅仅是判断两帧和音频帧是否能够对上，模型时逐帧处理的，误认为所有在音频中出现的声音也是说话的声音导致出现这种现象。

6 总结与展望

本文对Video-retalking进行了比较详细地介绍，同时在复现中实现了对D-net的训练并在本文提供了复现细节。由于本文还存在其他子网络，没来得及复现它们的训练是不足之处。在未来的研究中，我们可以考虑进一步扩展我们的工作，以包括这些子网络的训练和实现。

一种可能的方法是使用更多的数据集来训练这些子网络。Video-retalking是一个复杂的任务，涉及到许多不同的方面，如语音识别和计算机视觉等。因此，为了获得更好的性能，我们需要更多的数据来训练我们的模型。此外，我们还可以考虑在训练过程中加入一些新的技巧，如数据增强，正则化。

在复现过程中发现了模型的局限性：输出噪声后人物嘴依然在开合的问题上。提高唇语识别的准确性应该是Video-retalking任务的关键环节，准确识别唇语有助于生成更自然的语音，可以将模型改为非逐帧训练而、或者使用记忆单元训练来让模型识别噪声。或者直接对原音频训练一个降噪器，经过降噪机处理后音频在送入模型。

参考文献

[1] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019b. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *CVPR Workshops*.

[2] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. 2021. PIRenderer: Controllable Portrait Image Generation via Semantic Neural Rendering. In *ICCV*.

[3] Lu Chi, Borui Jiang, and Yadong Mu. 2020. Fast Fourier Convolution. In *NeurIPS*.

[4] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*.

[5] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*.

[6] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2018. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence* (2018).

[7] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. 2021c. Towards Real-World Blind Face Restoration with Generative Facial Prior. In *CVPR*.

[8] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *TOG* (2017).

[9] Xin Wen, Miao Wang, Christian Richardt, Ze-Yin Chen, and Shi-Min Hu. 2020. Photorealistic Audio-driven Video Portraits. *TVCG* (2020).

[10] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. 2020. Neural voice puppetry: Audio-driven facial reenactment. In *ECCV*.

[11] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 2020. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM Multimedia*.

[12] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. 2022. Everybody's talkin': Let me talk as you want. *IEEE Transactions on Information Forensics and Security* (2022).

[13] Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and CV Jawahar. 2019. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1428–1436.

[14] Joon Son Chung and Andrew Zisserman. 2016. Out of time: automated lip sync in the wild. In ACCV.

[15] Se Jin Park, Minsu Kim, Joanna Hong, Jeongsoo Choi, and Yong Man Ro. 2022. SyncTalkFace: Talking Face Generation with Precise Lip-syncing via Audio-Lip Memory. In 36th AAAI Conference on Artificial Intelligence (AAAI 22). Association for the Advancement of Artificial Intelligence.

[16] Tianyi Xie, Liucheng Liao, Cheng Bi, Benlai Tang, Xiang Yin, Jianfei Yang, Mingjie Wang, Jiali Yao, Yang Zhang, and Zejun Ma. 2021. Towards Realistic Visual Dubbing with Heterogeneous Sources. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1739–1747.

[17] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. 2022. Everybody's talkin': Let me talk as you want. *IEEE Transactions on Information Forensics and Security* (2022).

[18] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. 2020. Neural voice puppetry: Audio-driven facial reenactment. In *ECCV*.

[19] Yuanxun Lu, Jinxiang Chai, and Xun Cao. 2021. Live Speech Portraits: Real-Time Photorealistic Talking-Head Animation. *TOG* (2021).

[20] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. 2021. Audio-Driven Emotional Video Portraits. *arXiv preprint arXiv:2104.07452* (2021).

[21] Avisek Lahiri, Vivek Kwatra, Christian Frueh, John Lewis, and Chris Bregler. 2021. LipSync3D: Data-Efficient Learning of Personalized 3D Talking Faces from Video using Pose and Lighting Normalization. In *CVPR*.

[22] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*.

[23] Yang Song, Jingwen Zhu, Dawei Li, Xiaolong Wang, and Hairong Qi. 2018. Talking face generation by conditional recurrent adversarial network. *arXiv preprint arXiv:1804.04786* (2018).

[24] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. 2019. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI*.
[25] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2020. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision* 128, 5 (2020), 1398–1413.
[26] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7832–7841.
[27] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. 2020. MakeitTalk: speaker-aware talking-head animation. *TOG* (2020).
[28] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. 2021a. Flow-Guided One-Shot Talking Face Generation With a High-Resolution Audio-Visual Dataset. In *CVPR*.
[29] Suzhen Wang, Lincheng Li, Yu Ding, and Xin Yu. 2021a. One-shot Talking Face Generation from Single-speaker Audio-Visual Correlation Learning. *arXiv preprint arXiv:2112.02749* (2021).
[30] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First order motion model for image animation. *NIPS* (2019).
[31] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. 2021. PIRenderer: Controllable Portrait Image Generation via Semantic Neural Rendering. In *ICCV*.
[32] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*.
[33] Joon Son Chung and Andrew Zisserman. 2016. Out of time: automated lip sync in the wild. In *ACCV*.
[34] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. 2021. GAN Prior Embedded Network for Blind Face Restoration in the Wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[35] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. 2021c. Towards Real-World Blind Face Restoration with Generative Facial Prior. In *CVPR*.
[36] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019a. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*.
[37] Lu Chi, Borui Jiang, and Yadong Mu. 2020. Fast Fourier Convolution. In NeurIPS.