语音对口型专家是你所需要的一切，在野外进行语音对口型生成

K R Prajwal*
prajwal.k@research.iiit.ac.in
印度海得拉巴IIIT公司。

Rudrabha Mukhopadhyay*
radrabha.m@research.iiit.ac.in
IIIT, Hyderabad, India

Vinay P. Namboodiri
vpn22@bath.ac.uk 英国巴斯大学

C V Jawahar
jawahar@iiit.ac.in
IIIT，印度海得拉巴。

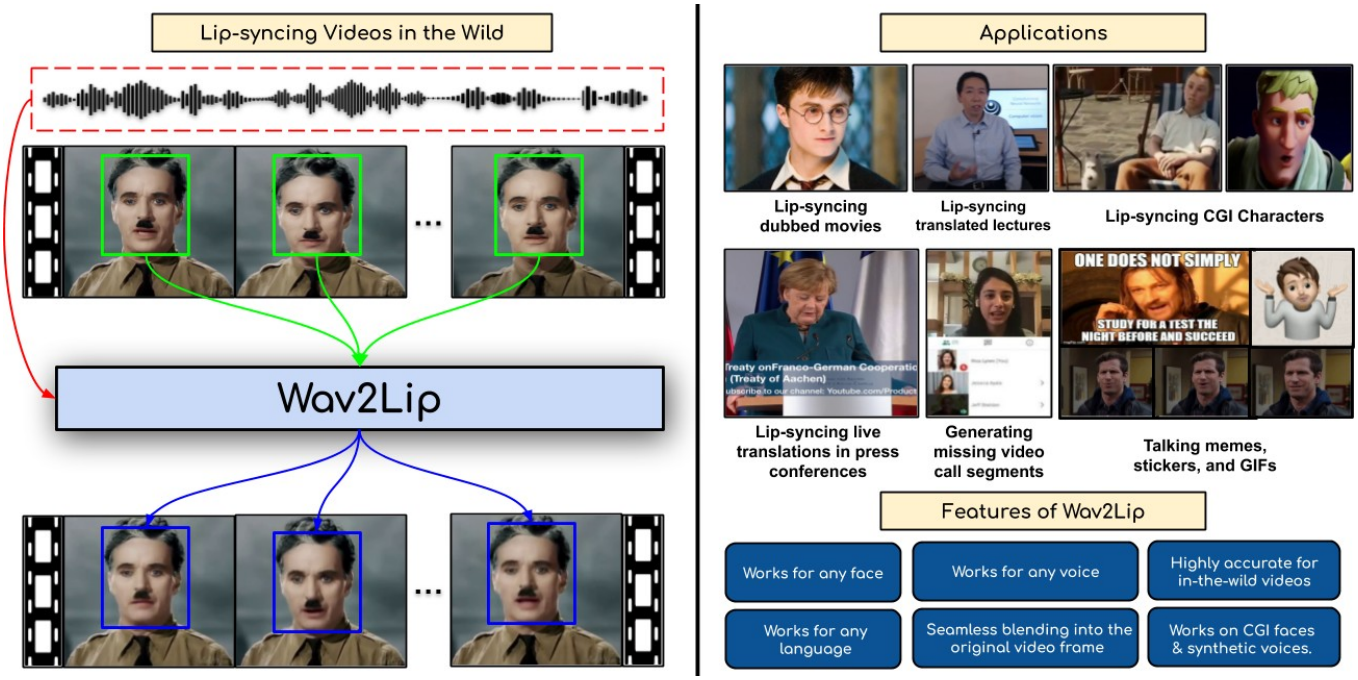

图1：我们的新型Wav2Lip模型在动态、无约束的人脸对话视频中产生了明显更准确的唇部同步。定量指标表明，我们生成的视频中的唇部同步几乎与真实同步的视频一样好。因此，我们相信我们的模型可以实现广泛的现实世界的应用，而以前的独立于说话人的唇语同步方法[17, 18]很难产生满意的结果。

ABSTRACT

在这项工作中，我们研究了对任意身份的人脸视频进行唇语同步以匹配目标语音片段的问题。目前的工作擅长在静态图像或训练期间看到的特定人物的视频上产生准确的唇部动作。

*两位作者对这项研究有同等贡献。

MM '20，2020年10月12-16日，美国华盛顿州西雅图。
ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413532

阶段。然而，在动态的、不受约束的人脸视频中，他们不能准确地变形任意身份的嘴唇运动，导致视频的很大一部分与新音频不同步。我们确定了与此相关的关键原因，并通过学习一个强大的唇部同步判别器来解决这些问题。接下来，我们提出了新的、严格的评估基准和指标，以准确地测量不紧张的视频中的唇部同步性。对我们具有挑战性的基准进行的广泛的定量评估表明，由我们的Wav2Lip模型生成的视频的唇部同步精度几乎与真实的同步视频一样好。我们在网站上提供了一个演示视频，清楚地显示了我们的Wav2Lip模型和评估基准的实质性影响：cvit.iiit.ac.in/research/projects/cvit-projects/a-lip-sync-expert-is-all-you-need-for-speech-to-lip-generation-in the-wild。代码和模型在这里发布：github.com/Rudrabha/Wav2Lip。你

也可以通过这个链接尝试互动演示：bhaasha.iiit.ac.in/ lipsync。

CCS概念

' → **计算方法计算机视觉; 从评论中学习**; *语音学/形态学*。

关键字

唇部同步;视频生成;说话表情生成

ACM参考格式。
K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C V Jawa- har.2020.唇部同步专家是你在野外生成语音到唇部所需要的一切。在第*28届ACM国际多媒体会议（MM'20）论文集，2020年10月12-16日，美国华盛顿州西雅图*。ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394171.3413532

1 简介

随着视听消费的指数式增长[21]，快速的视频内容创作已经成为一种典型的需求。同时，使这些视频能够以不同的语言访问也是一个关键的挑战。例如，深度学习系列讲座、著名电影或对国家的公开讲话，如果被翻译成所需的目标语言，就可以被数百万新的观众所接受。翻译这类人脸视频或创建新的视频的一个关键方面是纠正唇部同步以匹配所需的目标语音。因此，为匹配给定的输入音频流而对人脸视频进行唇语同步，在研究界受到了相当大的关注[6, 13, 17, 18, 23]。最初的工作[19，22]在这个空间里使用深度学习，利用单个说话人的七个小时学习了从语音表征到嘴唇地标的映射。最近的作品[13, 23]在这一路线上直接从语音表征中生成图像，并显示了他们所训练的特定演讲者的卓越的生成质量。然而，众多的实际应用需要能够随时用于一般身份和语音输入的模型。这导致了独立于说话人的语音到嘴唇生成模型的产生[17, 18]，这些模型是在数以千计的身份和声音上训练的。它们可以在任何身份、任何声音的单一静态图像上生成准确的唇部动作，包括由文本到语音系统生成的合成语音[18]。然而，为了用于翻译讲座/电视系列节目等应用，这些模型需要能够对这些动态的、不受约束的视频中存在的广泛的不同唇形进行变形。也是如此，而且不仅仅是在静态图像上。

我们的工作建立在这后一类独立于说话人的工作之上，这些工作希望与任何身份和声音的人脸视频进行唇语同步。我们发现，这些对静态图像效果很好的模型无法准确地对无约束的视频内容中的各种唇形进行变形，从而导致生成的视频中有很大一部分与新目标音频不同步。观众可以识别出一个不同步的视频片段，小到只有几秒钟。

≈ 0.05−0.1秒[9]的时间。因此，考虑到允许误差的微小程度，将真实世界的视频与全新的语音进行唇语对接是相当具有挑战性的。此外，我们的目标是一个独立于说话人的方法，没有任何额外的特定说话人的数据开销，这使得我们的任务更加复杂。

困难。真实世界的视频包含快速的姿势、比例和光照变化，生成的人脸结果也必须无缝地融合到原始目标视频中。

我们首先检查了现有的与说话人无关的语音到嘴唇的生成方法。我们发现，这些模型没有充分惩罚错误的唇形，这可能是由于只使用了重建损失或弱的唇同步判别器的结果。我们调整了一个强大的唇部同步判别器，它可以强制发生器持续产生准确、真实的唇部运动。接下来，我们重新审查了目前的评估协议，并设计了新的、严格的评估基准，这些基准来自三个标准测试集。我们还提出了可靠的评估指标，使用Sync- Net[9]来精确评估无约束视频中的唇部同步。我们还收集并发布了ReSyncED，这是一组具有挑战性的真实世界的视频，可以为模型在实践中的表现提供基准。我们进行了广泛的定量和主观的人类评估，并在所有的基准中以很大的优势超过了以前的方法。我们的主要贡献/主张如下。

- 我们提出了一种新型的唇语同步网络，即Wav2Lip，它比以前的作品在野外用任意的语音对任意的人脸视频进行唇语同步时的准确性要高得多。
- 我们提出了一个新的评估框架，由新的基准和指标组成，以便能够对无约束视频中的唇语同步进行公平的判断。
- 我们收集并发布了ReSyncED，这是一个真实世界的唇语同步评估数据集，用来衡量唇语同步模型在完全未见的野外视频上的表现。Wav2Lip是第一个独立于说话人的模型，它生成的视频的唇语同步精度与真实的同步视频相匹配。人类的评估表明，Wav2Lip生成的视频在90%以上的时间里比现有的方法和未同步的版本更受欢迎。

在我们的网站上可以找到一个演示视频[1]，其中有几个定性的例子，清楚地说明了我们模型的影响。我们还将在网站上发布一个互动演示，允许用户使用他们选择的音频和视频样本来尝试这个模型。本文的其余部分组织如下。第2节调查了语音转唇生成领域的最新发展，第3节讨论了现有作品的问题，并描述了我们提出的缓解这些问题的方法，第4节提出了一个新的、可靠的评估框架。我们在第5节中描述了各种潜在的应用，并解决了一些伦理方面的问题，在第6节中得出结论。

2 相关的工作

2.1 从语音中生成受限的会说话的脸

我们首先回顾一下关于说话脸谱生成的工作，这些工作要么是由于他们可以生成的身份范围有限，要么是由于他们的词汇范围有限。最近一些关于巴拉克-奥巴马视频的工作[19，22]实现了真实的人脸生成。他们在输入的音频和视频之间学习了一种映射

[1] cvit.iiit.ac.in/research/projects/cvit-projects/a-lip-sync-expert-is-all-you-need-for-speech to-lip-generation-in-the-wild

和相应的唇部地标。由于它们只对特定的说话人进行训练，因此它们不能为新的身份或声音进行合成。它们还需要大量的特定演讲者的数据，通常是几个小时。最近沿着这个思路的一项工作[13]提出通过添加或重新移动讲话中的短语来无缝编辑单个讲话者的视频。他们仍然需要每个演讲者一个小时的数据来实现这一任务。最近，另一项工作[23]试图通过使用一个两阶段的方法来最小化这种数据开销，他们首先学习与说话人无关的特征，然后用所需说话人的5分钟数据学习渲染映射。然而，他们在一个小得多的语料库上训练独立于说话人的网络，并且还有一个额外的开销，即需要每个目标说话人的干净训练数据来为该说话人生成。现有作品的另一个限制是在词汇方面。一些作品[5, 26, 28]以有限的词汇集进行训练，如GRID[10]（56个词）、TIMIT[14]和LRW[8]（1000个词），这大大阻碍了模型学习真实视频中大量的音素-词汇映射的多样性[18]。我们的工作重点是对无约束的人脸视频进行唇语匹配，以匹配任何目标语音，不受身份、声音或词汇的限制。

2.2 从语音中生成无约束的会说话的人脸

尽管在语音驱动的人脸生成方面的工作越来越多，但令人惊讶的是，很少有工作被设计成与任意身份、声音和语言的视频对口。它们没有在一小部分身份或一小部分词汇上进行训练。这使得它们在测试时可以对任何语音的随机身份进行唇语同步。据我们所知，在目前的文献中，只有两个这样突出的作品[17, 18]存在。请注意，[17]是[7]的扩展版本。这两项工作[17，18]都将在野外学习唇语的任务表述如下。*给定一个简短的语音片段S和一个随机的参考人脸图像R，网络的任务是生成一个与音频相匹配的输入人脸的唇语版本*L_{∂}。此外，LipGAN模型还输入了下半身被遮住的目标脸，作为姿势的先验。这一点至关重要，因为它允许将生成的人脸作物无缝粘贴到原始视频中，而无需进一步的后期处理。它还与生成器一起训练了一个鉴别器，以鉴别同步或不同步的音视频对。然而，这两项工作都有一个显著的局限性：它们在任意身份的静态图像上工作得非常好，但在试图对野外无约束的视频进行唇语生成时，却产生了不准确的唇语。与LipGAN[18]中使用的GAN设置不同，我们使用了一个预先训练好的、准确的唇语识别器，该识别器没有与生成器一起进一步训练。我们观察到，这是一个重要的设计选择，可以实现更好的唇语结果。

3 准确的语音驱动的野外视频对口型

我们的核心架构可以被概括为 *"通过向训练有素的唇语专家学习来生成准确的唇语"*。为了理解这一设计选择，我们首先确定了现有架构（第2.2节）在野外视频中产生不准确的唇语的两个关键原因。我们认为，损失函数，即

现有作品[17, 18]中使用的L1重建损失和LipGAN[18]中的判别器损失都不足以惩罚不准确的唇语生成。

3.1 像素级重建损失是判断对口型的弱点

脸部重建损失是针对整个图像计算的，以确保正确的姿势生成，保留身份，甚至脸部周围的背景。嘴唇区域相当于总重建损失的4%以下（基于空间范围），因此在网络开始执行细粒度的嘴唇形状校正之前，大量的周围图像重建首先被优化。网络在训练过程中（20个历时[18]）的一半左右（*第11个历时*）才开始对嘴唇进行变形，这一事实进一步证明了这一点。因此，有一个额外的判别器来判断唇部同步是至关重要的，LipGAN[18]也是如此。但是，LipGAN中采用的判别器有多强大？

3.2 一个弱的唇音辨别器

我们发现，LipGAN的唇部同步判别器在LRS2测试集上检测不同步的音频-唇部对时，准确率只有56%左右。作为比较，我们将在这项工作中使用的专家鉴别器在相同的测试集上有91%的准确性。我们假设这种差异有两个主要原因。首先，LipGAN的判别器使用单帧来检查唇语同步。在表3中，我们表明，在检测唇语同步时，小的时间背景非常有帮助。其次，在训练过程中生成的图像由于尺度和姿势变化较大而含有大量的伪影。我们认为，像LipGAN那样，在GAN设置中对这些嘈杂的生成图像进行训练，会导致鉴别器专注于视觉伪影而不是音频-嘴唇的对应关系。这导致了非同步检测精度的大幅下降（表3）。我们认为并表明，从实际视频帧中捕捉到的 "真实"、准确的唇部同步概念可用于准确判别和执行生成图像中的唇部同步。

3.3 你只需要一个对口型专家就够了

基于以上两个发现，我们建议使用一个预先训练好的专家级唇语判别器，它能准确地检测出真实视频中的同步性。此外，它不应该像LipGAN那样在生成的帧上进一步微调。SyncNet[9]模型就是这样的一个网络，它被用来纠正用于创建大型唇部同步数据集的唇部同步错误[1, 3]。我们建议为我们的任务调整和训练SyncNet[9]的修改版本。

*3.3.1 SyncNet的概述。*SyncNet[9]输入一个由T_v个连续人脸帧（仅下半部分）组成的窗口V和一个大小为T_a D的语音段S，其中T_v和T_a分别是视频和音频的时间步长。它被训练成可以通过随机取样音频窗口T_a D来区分音频和视频的同步性，该窗口要么与视频对齐（同步），要么来自不同的时间步长（不同步）。它包含一个人脸编码器和一个音频编码器，两者都是由一叠二维转折组成的。从这些编码器生成的嵌入之间的L2距离被计算出来，模型被训练成最大边际损失，以最小化（或最大化）同步（或不同步）对之间的距离。

≈ ≈

× ×

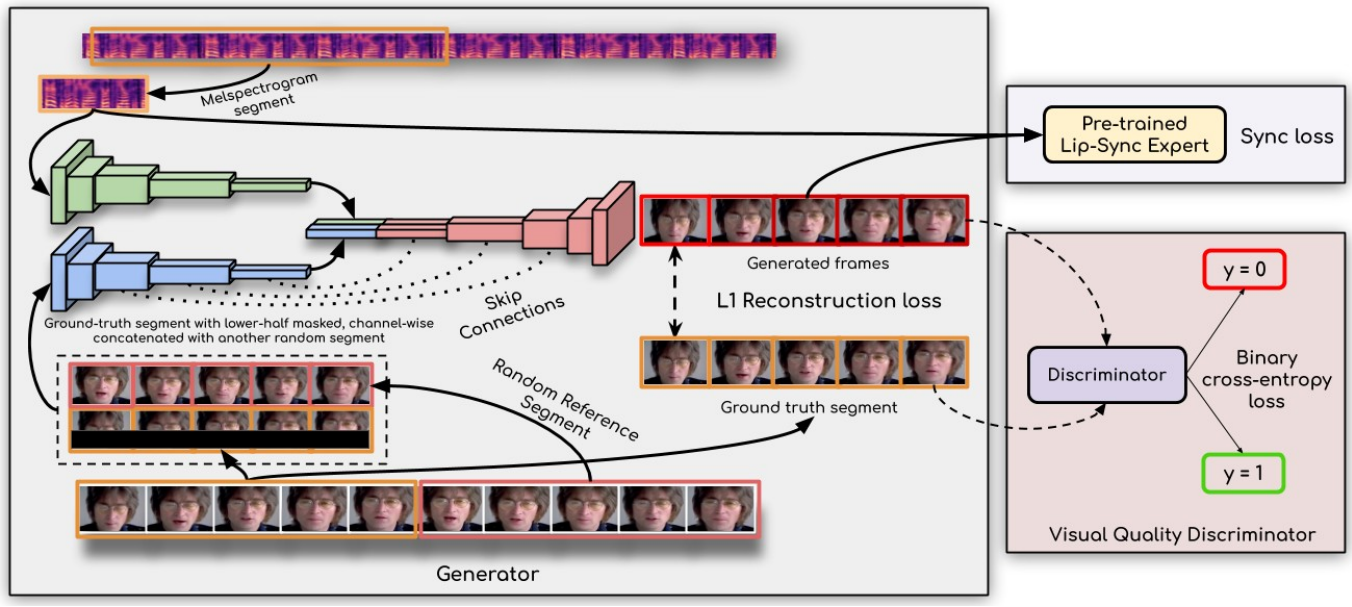

图2：我们的方法通过向 "已经训练有素的唇语专家 "学习，产生准确的唇语。与以往只采用重建损失[17]或在GAN设置中训练判别器[18]的工作不同，我们使用了一个预先训练好的判别器，它在检测唇语错误方面已经相当准确。我们表明，在嘈杂的生成面孔上进一步微调它，会妨碍判别器测量唇部同步的能力，从而也会影响生成的唇形。此外，我们还采用了一个视觉质量鉴别器来提高视觉质量和同步的准确性。

*3.3.2 我们的专家级唇语鉴别器。*我们对SyncNet[9]做了以下修改，以训练一个适合我们唇语生成任务的专家级唇语识别器。首先，我们不再像原始模型那样提供灰度图像，而是提供彩色图像。其次，我们的模型明显更深，有剩余的跳过连接[15]。第三，受这个公共实现[2]的启发，我们使用了一个不同的损失函数：余弦相似性与二进制交叉熵损失。也就是说，我们计算ReLU激活的视频和语音嵌入v，s之间的点积，为每个样本产生一个介于0，1之间的单一值，表明输入的音频视频对是同步的概率。

$$\text{Psync} = \frac{v - s}{max(\|v\|_2 - \|s\|_2, \epsilon)} \quad (1)$$

我们在LRS2训练中训练我们的专家级唇语辨别器分割（29小时），批次大小为64，T_v=5帧，使用亚当优化器[12]，初始学习率为$1e{-3}$。我们的专家的唇语鉴别器在LRS2测试集上的准确率约为91%，而LipGAN中使用的鉴别器在同一测试集上的准确率仅为56%。

3.4 通过向对口型专家学习生成准确的对口型

现在我们有了一个准确的唇语辨别器，我们现在可以用它来惩罚发生器（图2）在训练期间不准确的生成。我们首先描述一下发生器的结构。

[2] github.com/joonson/syncnet_trainer

*3.4.1 生成器架构细节。*我们使用与LipGAN[18]类似的发生器架构。我们的关键贡献在于用专家判别器来训练它。生成器G包括三个模块。(i) 身份编码器，(ii) 语音编码器，和(iii) 面部解码器。身份编码器是一堆残余卷积层，对随机参考帧R进行编码，沿通道轴与姿势先验P（下半身被遮住的目标脸）相联系。语音编码器也是一个二维卷积层，对输入的语音片段S进行编码，然后将其与人脸表示相连接。解码器也是一个卷积层的堆叠，同时还有用于上采样的转置卷积。生成器被训练成最小的L1重建产生的框架L_G和地面真实框架L_G之间的损失。

$$L_{\text{recon}} = \frac{1}{N}\sum_{i=1}^{N} \|L_G - L_G\|_1 \quad (2)$$

因此，生成器与之前的工作类似，是一个2D-CNN编码器-解码器网络，独立生成每一帧。那么，我们如何使用我们预先训练好的专家唇语辨别器，它需要一个T_v=5帧的时间窗口作为输入？

*3.4.2 惩罚不准确的嘴唇生成。*在训练过程中，由于第3.3节中训练的专家鉴别器一次处理T_v=5个连续帧，我们还需要生成器G来生成所有T_v=5的帧。我们为参考帧取样一个随机的连续窗口，以确保整个T_v窗口内姿势等的时间一致性。由于我们的发生器独立地处理每一帧，我们将时间步数沿以下方向堆叠

在送入参考帧的同时，得到一个$N_{Tv}, H, W, 3$的输入形状，其中N, H, W分别是批次大小、高度和宽度。在将生成的帧送入专家鉴别器时，时间步长沿通道维度串联，这在鉴别器的训练中也是如此。专家判别器的输入形状是$N, H/2, W, 3_{Tv}$，其中只有生成的脸的下半部分被用于判别。生成器也被训练成最小化来自专家鉴别器的"专家同步损失"$_{\text{Esync}}$。

$$\text{Esync} = \frac{1}{N}\sum_{i=1}^{N} -\log(P_{\text{siync}}) \tag{3}$$

其中$_{\text{Psiync}}$是根据公式1计算的。请注意，专家判别器的权重在发生器的训练过程中保持冻结。这种纯粹基于的强有力的判别从真实视频中学习到的唇部同步概念迫使生成器也要实现真实的唇部同步，以尽量减少唇部同步损失$_{\text{Esync}}$。

3.5 生成照片般真实的面孔

在我们的实验中，我们观察到，使用一个强大的唇部同步干扰器迫使发生器产生准确的唇形。然而，它有时会导致变形的区域略微模糊或包含轻微的人工痕迹。为了减轻这种轻微的质量损失，我们在GAN设置中与生成器一起训练一个简单的视觉质量判别器。因此，我们有两个鉴别器，一个用于同步准确性，另一个用于更好的视觉质量。由于3.2中解释的原因，唇语同步判别器*没有*在GAN设置中训练。另一方面，由于视觉质量判别器不对唇语同步进行任何检查，只对不真实的人脸生成进行惩罚，所以它是在生成的人脸上训练的。

鉴别器D由一叠卷积块组成。每个块由一个卷积层和一个Leaky ReLU激活组成[20]。鉴别器被训练为最大化目标函数L_{disc}（公式5）。

$$L_{\partial en} = E_{x\sim L_{\partial}}[log\partial(1-D(x)] \tag{4}$$

$$L_{disc} = E_{x\sim L_G}[log\partial(D(x))] + L_{\partial en} \tag{5}$$

其中L_{∂}对应于来自发生器G的图像，而$_{LG}$与真实图像相对应。

生成器最小化方程6，它是重建损失（方程2）、同步损失（方程3）和对抗性损失$L_{\partial en}$（方程4）的加权和。

$$\text{Ltotal} = (1 - s_w - s_{\partial}) \cdot \text{Lrecon} + s_w \cdot \text{Esync} + s_{\partial} \cdot L_{\partial en} \tag{6}$$

其中，$_{sw}$是同步惩罚权重，s_{∂}是广告。

在我们所有的实验中，经验性地设定为0.03和0.07的精算损失。因此，我们的完整网络使用两个互不相干的判别器对同步准确率和质量进行了优化。我们只在LRS2训练集[1]上训练我们的模型

它是如何在真实视频的推理过程中工作的。与Lip- GAN[18]类似，该模型逐帧地生成一个会说话的人脸视频。每个时间步的视觉输入是当前的人脸裁剪（来自源帧），并与相同的当前人脸裁剪相连接，其下半部分被屏蔽，作为姿势先验。因此，在推理过程中，模型不需要改变姿势，大大减少了伪影。相应的音频片段也被作为输入给语音子网络，该网络生成输入的面部裁剪，但嘴部区域被变形。

我们所有的代码和模型都将公开发布。我们现在将对照以前的模型，定量评估我们的新方法。

4 定量评价

尽管只在LRS2训练集上进行训练，我们还是在3个不同的数据集上评估了我们的模型。但在这之前，我们重新调查了之前工作中所遵循的当前评估框架，以及为什么它远不是评估这一领域工作的理想方式。

4.1 重新思考语音驱动的野外对口相声的评估框架

目前独立于说话人的唇语同步的评估框架对模型的判断与对真实视频进行唇语同步的方式不同。具体来说，不是将当前帧作为参考（如上一节所述），而是选择视频中的一个主帧作为参考，以便在评估时不泄露正确的唇语信息。我们强烈认为，前一段的评估框架对于评估唇语同步的质量和准确性并不理想。经过对上述评估系统的仔细研究，我们观察到一些关键的局限性，我们在下面讨论。

*4.1.1 不能反映真实世界的使用情况。*如前所述，在测试时的生成过程中，模型必须不改变姿势，因为生成的人脸需要无缝粘贴到框架中。然而，目前的评估框架在输入中提供了随机的参考框架，因此要求网络改变姿势。因此，上述系统并不能评估模型在现实世界中的使用情况。

*4.1.2 不一致的评价。*由于参考框架是随机选择的，这意味着不同作品的测试数据是不一致的。这将导致不公平的比较，阻碍结果的可重复性。

*4.1.3 不支持检查时间上的一致性。*由于参考框架是在每个时间步长中随机选择的，所以时间上的

由于帧是随机生成的，所以已经失去了一致性。
姿势和规模。目前的框架不能支持一个新的元

，批次大小为80。我们使用亚当优化器[12]，其初始学习速率为$^{1e-4}$，betasβ_1=0.5，β_2=0.999，对发电机都是如此。
请注意，唇部同步判别器没有被进一步微调，所以它的权重被冻结。我们通过解释以下内容来结束对我们所提出的架构的描述

这是个很好的方法，旨在研究这个问题的时间一致性方面。

*4.1.4 目前的衡量标准并不是专门针对唇语同步的。*现有的指标，如SSIM[27]和PSNR，是为了评估所有的图像质量而不是细粒度的唇部同步错误。尽管LMD[4]专注于唇部区域，但我们发现在生成的人脸上，唇部地标可能是相当不准确的。因此，需要有一个专门用于测量唇部同步误差的指标。

	LRW [8]			LRS2 [1]			LRS3 [3]		
方法	**LSE-D ↓**	**LSE-C ↑**	**FID ↓**	**LSE-D ↓**	**LSE-C ↑**	**FID ↓**	**LSE-D ↓**	**LSE-C ↑**	**FID ↓**
Speech2Vid [17]	13.14	1.762	11.15	14.23	1.587	12.32	13.97	1.681	11.91
LipGAN[18]	10.05	3.350	2.833	10.33	3.199	4.861	10.65	3.193	4.732
Wav2Lip (我们的)	**6.512**	**7.490**	3.189	**6.386**	**7.789**	4.887	**6.652**	**7.887**	4.844
Wav2Lip + GAN (我们的)	6.774	7.263	**2.475**	6.469	7.781	**4.446**	6.986	7.574	**4.350**
真实视频	7.012	6.931	-	6.736	7.838	-	6.956	7.592	-

表1：我们提出了两个新的指标 "Lip-Sync Error-Distance"（越低越好）和 "Lip-Sync Error-Confidence"（越高越好），它们可以可靠地衡量无约束视频的唇部同步准确性。我们看到，使用Wav2Lip生成的视频的唇部同步准确性几乎与真实的同步视频一样好。请注意，我们只在LRS2[1]的训练集上进行了训练，但我们在所有的数据集上都能轻松地进行推广，而不需要进一步的微调。我们还报告了FID得分（越低越好），这清楚地表明，使用视觉质量判别器可以显著提高质量。

4.2 评价野外对口型的新基准和衡量标准

对随机帧进行采样评估的原因是，当前帧已经与语音同步，导致输入本身的唇形泄漏。而以前的工作没有尝试对不同的语音片段进行采样，而不是对不同的帧进行采样，因为采样的语音的基础真实唇形是不可用的。

*4.2.1 衡量唇部同步误差的指标。*我们建议使用预先训练好的SyncNet[9]来测量生成的帧和随机选择的语音段之间的唇部同步误差。SyncNet在一个视频片段中的平均准确率超过99%[9]。因此，我们相信这可以成为一个很好的自动评估方法，明确地测试野外无约束视频中的准确唇语同步。请注意，这不是我们上面训练的专家级唇语鉴别器，而是Chung和Zisserman[9]发布的，它是在一个不同的、非公开的数据集上训练的。使用SyncNet解决了现有评估框架的主要问题。我们不再需要对随机的、时间上不连贯的帧进行采样，而且SyncNet在评估唇语同步时还考虑到了短距离的时间一致性。因此，我们提出了两个使用SyncNet模型自动确定的新指标。第一个是以唇部和音频表示之间的距离计算的平均误差指标，我们将其命名为 "LSE-D"（"唇部同步误差-距离"）。较低的LSE-D表示较高的视听匹配度，即语音和嘴唇运动是同步的。第二个指标是平均信心分数，我们将其命名为 "LSE-C"（唇部同步误差-距离）。
- 信心）。信心越高，音频视频就越好
相关性。一个较低的置信度分数表示视频中有几个部分的唇部动作完全不同步。进一步的细节可以在SyncNet论文[9]中找到。

*4.2.2 一个一致的基准来评估野外的唇语同步。*现在，我们有了一个可以为任何视频和音频对计算的自动、可靠的指标，我们可以在每个时间步长中随机抽取语音样本而不是随机帧。因此，我们可以创建一个视频和伪随机选择的音频对的列表，作为一个一致的测试集。我们创建了三个一致的基准测试集，分别使用LRS2[1]、LRW[8]和LRS3[3]的测试集视频各一个。对于每个视频V_S，我们从另一个随机抽样的视频V_I中抽取音频，条件是V_I的长度为1.5米。

[3] github.com/joonson/syncnet_python

语音的V_I要小于V_S。我们使用LRS2创建了*14K*音频-视频对。使用LRW测试集，我们创建了*28K*对，这个测试集衡量了正面/近正面视频的性能[2]。我们还使用LRS3测试集创建了*14K*个对，这也将是轮廓视图中对口型的基准。完整的评估工具包将被公开发布，以便在野外对对口型视频进行一致和可靠的基准测试。

4.3 在新基准上比较各种模型

我们使用LSE-D和LSE-C指标在我们新创建的测试集上比较前两种方法[17, 18]。在进行测试的过程中，我们现在在每个时间步中提供相同的参考和姿势优先，这与之前在结构部分的描述相似。表1显示了所有三个测试中的音频-视频对的平均LSE-D和LSE-C得分。此外，为了衡量生成的面孔的质量，我们还报告了Frél'chet Inception Distance（FID）。我们的方法在很大程度上超过了以前的方法，这表明了强大的唇语识别能力的重要作用。我们还可以看到，在使用视觉质量鉴别器和唇音专家鉴别器之后，质量有了明显的改善。然而，我们观察到在使用视觉质量判别器后，同步的准确性略有下降。因此，我们将发布这两个模型，因为它们在视觉质量和同步准确性之间有一个轻微的权衡。

4.4 真实世界的评估

除了在标准数据集上进行评估外，我们新的评估框架和指标允许我们在真实世界的视频上进行评估，这些模型最有可能被使用。此外，鉴于人类对音唇同步的敏感性[9]，有必要在人类评价者的帮助下评估我们的结果。因此，与之前的独立于说话者的唇语工作相反，我们首次在网络上无约束的真实视频上进行了定量和人的评估实验。因此，我们收集并公开发布了 "ReSyncED " 这一 "真实世界评估数据集"，以主观和客观地衡量唇语同步工作的性能。

*4.4.1 策划ReSyncED。*我们所有的视频都是从YouTube下载的。我们特别选择了三种类型的视频例子。第一种类型 "配音"，包含音频自然不同步的视频，如配音的电影片段或被现场翻译成不同

语言的公共演讲（因此演讲者的嘴唇

方法	视频类型	LSE-D ↓	LSE-C ↑	FID ↓	同步速度。	视觉素质。	总体经验	倾向性
不同步的原件。视频	配音	12.63	0.896	-	0.21	4.81	3.07	3.15%
Speech2Vid [17]		14.76	1.121	19.31	1.14	0.93	0.84	0.00%
LipGAN[18]		10.61	2.857	12.87	2.98	3.91	3.45	2.35%
Wav2Lip (我们的)		**6.843**	**7.265**	15.65	**4.13**	3.87	4.04	34.3%
Wav2Lip + GAN (我们的)		7.318	6.851	**11.84**	4.08	**4.12**	**4.13**	**60.2%**
没有对口型的	随机	17.12	2.014	-	0.15	4.56	2.98	3.24%
Speech2Vid [17]		15.22	1.086	19.98	0.87	0.79	0.73	0.00%
LipGAN[18]		11.01	3.341	14.60	3.42	3.77	3.57	3.16%
Wav2Lip (我们的)		**6.691**	**8.220**	14.47	**4.24**	3.68	4.01	29.1%
Wav2Lip + GAN (我们的)		7.066	8.011	**13.12**	4.18	**4.05**	**4.15**	**64.5%**
没有对口型的	TTS	16.89	2.557	-	0.11	4.67	3.32	8.32%
Speech2Vid [17]		14.39	1.471	17.96	0.76	0.71	0.69	0.00%
LipGAN[18]		10.90	3.279	11.91	2.87	3.69	3.14	1.64%
Wav2Lip（我们的）		**6.659**	**8.126**	12.77	**3.98**	3.87	3.92	41.2%
Wav2Lip + GAN（我们的）		7.225	7.651	**11.15**	3.85	**4.13**	**4.05**	**51.2%**
未翻译的视频		7.767	7.047	-	4.83	4.91	-	-

表2:使用我们新收集的ReSyncED基准进行的真实世界评估。我们使用定量指标和人类评价分数对三类真实视频进行评估。我们可以看到，在所有情况下，Wav2Lip模型都能产生高质量、准确的对口型视频。具体来说，指标表明我们的唇语同步视频和真实的同步视频一样好。我们还注意到，人类的评价表明，在试图对TTS生成的语音进行唇语同步时，还有一个改进的余地。最后，值得注意的是，在90%以上的时间里，我们的唇语视频比现有的方法或实际未同步的视频更受欢迎。

是与翻译的语音不同步的）。第二种类型是 "随机"，我们有一个视频集合，我们创建类似于4.2.2的随机视听对。第三种也是最后一种类型的视频，即 "TTS"，是专门为测试从文本到语音系统获得的合成语音的唇语性能而选择的。这对于未来有志于自动翻译视频（面对面翻译[18]）或快速创建新视频内容的工作至关重要。我们手动转录文本，使用谷歌翻译（总共约5种语言）和公开可用的文本到语音模型，为该类别的视频生成合成翻译语音。我们的任务是纠正原始视频中的唇部动作，以匹配这个合成语音。

*4.4.2 ReSyncED的真实世界评估。*我们首先使用从SyncNet[9]获得的新的自动指标 "LSE-D "和 "LSE-C "评估生成的真实视频结果。对于人工评估，我们要求14位评估员根据以下参数来判断不同的同步版本的视频。(a) 同步精度 (b) 视觉质量（评估视觉伪影的程度），(c) 整体体验（评估视听内容的整体体验），以及(d) 偏好，即观众选择最吸引人的视频版本观看。前三个参数的得分在1 5之间，(d)是单选投票，我们报告一个模型获得的投票百分比。我们分别评估三类视频中的每一类，并在表2中报告我们的结果。一个值得注意的结果是，以前的工作[17，18]产生了几个不同步的片段，但与不同步的版本相比，不那么受欢迎，因为后者仍然保留了良好的视觉质量。因此，我们的工作是第一个在野外对未同步的人脸视频进行了重大改进。我们还在图3中显示了一些定性的比较，其中包含了一些从ReSyncED测试集生成的样本。

4.5 我们的专家鉴别器在备选方案中是最好的吗？

模型	微调？	脱离同步加速。	LSE-D	LSE-C
T_V = 1 [18]	✓	55.6%	10.33	3.19
我们的，T_V = 1	×	79.3%	8.583	4.845
我们的，T_V = 3	✓	72.3%	10.14	3.214
我们的，T_V = 3	×	87.4%	7.230	6.533
我们的，T_V=5	✓	73.6%	9.953	3.508
我们的，T_V=5	×	**91.6%**	**6.386**	**7.789**

表3：更大的时间窗口可以更好地分辨唇语。另一方面，在生成的面孔上训练唇部同步判别器会降低其检测不同步的音频-嘴唇对的能力。因此，使用这样的判别器来训练唇语生成器，会导致唇语同步性差的视频。

我们的专家判别器使用T_V=5个视频帧来测量唇语误差。在GAN设置中，它也没有对生成的面孔进行微调。我们在这个消融研究中证明了这两个设计选择的合理性。我们可以通过从LRS2测试集中随机抽取同步和非同步对来测试判别器的性能。我们改变T_V=1、3、5的大小以了解其对检测同步的影响。在训练Wav2Lip模型时，我们还对T_V的三个变体进行了微调/冻结。因此，我们在表3中共得到了6种变体，从中我们可以清楚地看到两种情况。增加时间窗口的大小T_V始终能提供更好的唇语识别性能。更重要的是，我们看到，如果我们对生成的含有假象的脸部进行微调，那么鉴别器就会失去检测不同步的视听对的能力。我们认为，这种情况的发生是因为微调后的鉴别器将注意力集中在生成的人脸中的视觉假象上。

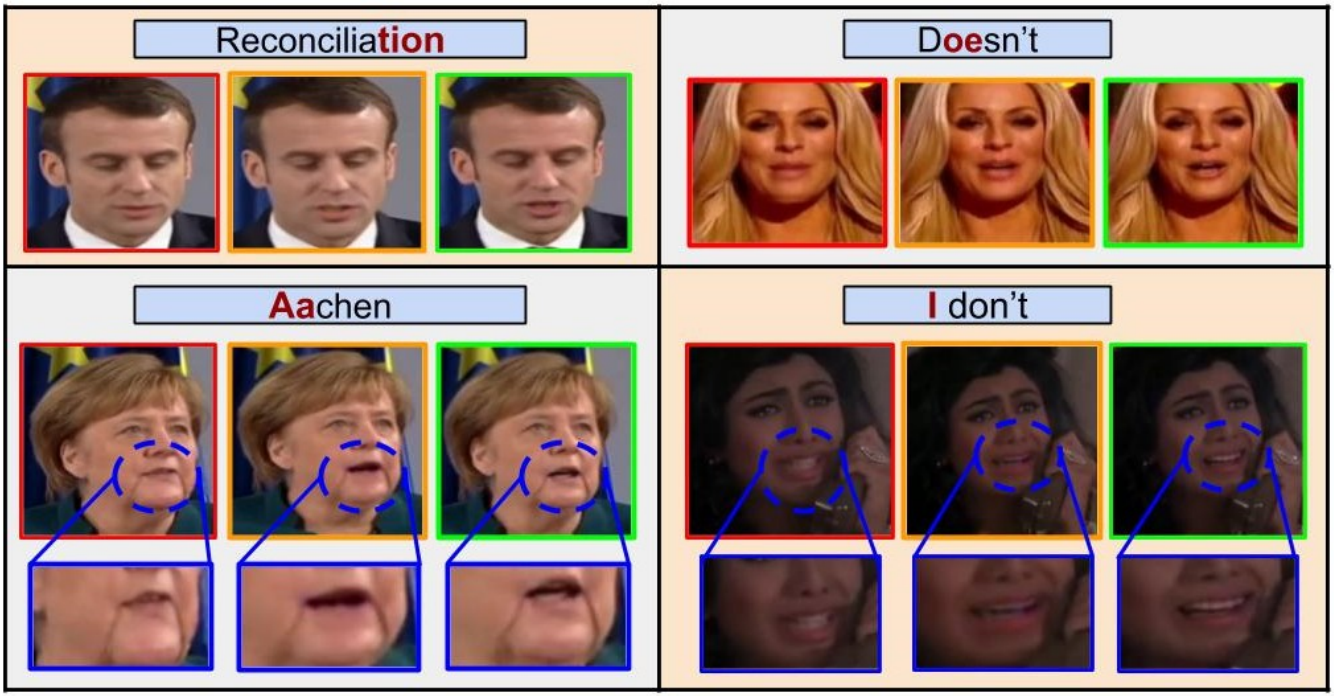

图3：由我们提出的模型生成的人脸实例（绿色和黄色的轮廓）。我们与目前最好的方法[18]（红色轮廓）进行比较。图中的文字是为了说明问题，以表示在图中的框架中正在说的话语。我们可以看到，我们的模型产生了准确、自然的唇形。增加一个视觉质量判别器也大大改善了视觉质量。我们强烈鼓励读者查看我们网站上的演示视频。

而不是精细的音频-嘴唇的反应。因此，它将*真正的*未同步的对子归类为 "同步"，因为这些真实的脸部图像不包含任何伪影。从长远来看，使用这样一个弱的判别器会导致我们的生成器对口型的惩罚不力，从而导致口型不对的人脸视频。

5 应用和合理使用

在我们的内容消费和社会交流越来越趋向于视听化的时候，迫切需要大规模的视频翻译和创作。Wav2Lip可以在满足这些需求方面发挥重要作用，因为它对野外的视频是准确的。例如，通常是英语的在线讲座视频现在可以与其他当地语言的（自动）配音语音进行唇语同步（表2，最后一个区块）。我们还可以为配音的电影配上唇语，使其成为令人愉快的观看方式（表2，第一部分）。在全球范围内，每天都有新闻发布会和公共讲话被现场翻译，但讲话者的嘴唇与转译的讲话不同步。我们的模型可以无缝地纠正这一点。在制作动画电影和丰富的对话式游戏内容时，将CGI角色的嘴唇与配音演员的讲话自动制成动画，可以节省几个小时的手工劳动。我们在网站上的演示视频中展示了我们的模型在所有这些应用和更多的应用。

我们认为，讨论和促进对能力越来越强的唇语作品的公平使用也是非常重要的。唇语的广泛适用性

我们的模型对任何身份和声音都具有近乎真实的唇语能力，这引起了人们对滥用可能性的关注。因此，我们强烈建议，使用我们的代码和模型创建的任何结果都必须明确地表明自己是合成的。除了上述强烈的积极影响外，我们将我们的工作完全开源的意图是，它可以同时也鼓励在检测被操纵的视频内容及其滥用方面的努力[11, 16, 24, 25]。我们相信，Wav2Lip可以实现七种积极的应用，也可以鼓励有关公平使用合成内容的富有成效的讨论和研究工作。

6 结论

在这项工作中，我们提出了一种新的方法，在野外生成准确的唇语同步视频。我们强调了目前的方法在对不紧张的人脸视频进行唇语同步时不准确的两个主要原因。在此基础上，我们认为预先训练好的、准确的唇部同步 "专家 "可以强制执行准确、自然的唇部动作生成。在评估我们的模型之前，我们重新审视了当前的定量评估框架，并强调了几个主要问题。为了解决这些问题，我们提出了几个新的评价基准和指标，还提出了一个真实世界的评价集。我们相信未来的工作可以在这个新框架中得到可靠的评判。我们的Wav2Lip模型在定量指标和人类评价方面都比目前的方法有很大的优势。我们还在一项消融研究中调查了我们对消融器的设计选择背后的原因。我们鼓励读者观看我们网站上的演示视频。

我们相信我们的努力和想法

在这个问题上的研究可以带来新的方向，例如在准确的唇部运动的同时合成表情和头部姿势。

参考文献

[1] T.Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman.2018.深度音频-视觉语音识别。在*arXiv:1809.02108*。

[2] T.Afouras, J. S. Chung, and A. Zisserman.2018.The Conversation:深度音频-视觉语音增强。在*INTERSPEECH*。

[3] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman.2018.LRS3-TED: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*（2018）.

[4] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu.2018. 唇部动作生成一目了然。在*欧洲计算机视觉会议（ECCV）论文集*。520-535.

[5] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu.2019.具有动态像素级损失的分层跨模式说话人脸生成。在*IEEE计算机视觉和模式识别会议的论文集中*。7832-7841.

[6] Lele Chen, Haitian Zheng, Ross K Maddox, Zhiyao Duan, and Chenliang Xu.2019.从声音到视觉：层次化的跨模式说话的脸部视频生成。在*IEEE计算机协会的计算机视觉和模式识别研讨会上*。

[7] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman.2017.你说什么？*arXiv预印本arXiv:1705.02966*（2017）。

[8] Joon Son Chung和Andrew Zisserman。2016.野外的读唇术。In *Asian Conference on Computer Vision*.Springer, 87-103.

[9] Joon Son Chung和Andrew Zisserman。2016.超时空：野外的自动唇语同步。在*多视角读唇术研讨会上，ACCV*。

[10] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao.2006.一个用于语音感知和自动语音识别的视听语料库。*美国声学学会杂志》*120，5（2006），2421-2424。

[11] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer.2020.DeepFake检测挑战数据集。arXiv:2006.07397 [cs.CV]

[12] John Duchi, Elad Hazan, and Yoram Singer.2011.用于在线学习和随机优化的自适应子梯度方法.*机器学习研究杂志》*12, 7 (2011).

[13] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala.2019.基于文本的说话头视频编辑。*ACM Transactions on Graphics（TOG）*38，4（2019），1-14。

[14] Naomi Harte和Eoin Gillen。2015.TCD-TIMIT:一个连续语音的视听语料库。*IEEE Transactions on Multimedia* 17, 5 (2015), 603-615.

[15] 何开明，张翔宇，任少卿，和孙健。2016.用于图像识别的深度残差学习。在*IEEE计算机视觉和模式识别会议论文集中*。770-778.

[16] 徐志忠，庄一秀，和李佳燕。2020.基于配对学习的深度假象检测.*应用科学*10 (2020), 370.

[17] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman.2019.你说什么？从音频中合成说话的脸。*International Journal of Computer Vision* 127，11-12 (2019), 1767-1779.

[18] Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Nam- boodiri, and CV Jawahar.2019.迈向自动面对面的翻译。In *Proceedings of the 27th ACM International Conference on Multimedia*.ACM, 1428-1436.

[19] Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre de Brébisson, and Yoshua Bengio.2017.Obamanet。*arXiv预印本arXiv:1801.01442*（2017）。

[20] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng.2013.整流器非线性改善神经网络声学模型。In *Proc. icml*, Vol. 30.3.

[21] NPD.2016.https://www.npd.com/wps/portal/npd/us/ news/press-releases/2016/52-percent-of-millennial-smartphone-owners-use- their-device for-video-calling-according-to-the-npd-group/，*52%的千禧年智能手机用户使用其设备进行视频通话*。

[22] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman.2017.合成奥巴马：从音频中学习唇部同步。*ACM Transactions on Graphics（TOG）*36，4（2017），95。

[23] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner.2019.Neural Voice Puppetry:*arXiv preprint arXiv:1912.05566* (2019). 音频驱动的面部重演。

[24] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia.2020.DeepFakes and Beyond:a Survey of Face Manipulation and Fake Detection. arXiv:2001.00179 [cs.CV]

[25] 埃莉诺-图斯曼、玛丽莲-乔治、塞尼-卡马拉和詹姆斯-汤普金。2020.Towards Untrusted Social Video Verification to Combat Deepfakes via

Face Geometry Consistency.在*IEEE/CVF 计算机视觉和模式识别会议（CVPR）研讨会论文集*。

[26] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic.2019.使用甘斯的真实语音驱动的面部动画。*International Journal of Computer Vision* (2019), 1-16.

[27] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. 2004.图像质量评估：从错误可见性到结构相似性。*IEEE transactions on image processing* 13, 4 (2004), 600-612.

[28] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang.2018.*arXiv preprint arXiv:1807.07860* (2018).