

***输入图像***

**学习现实的三维运动系数为样式化的音频驱动的单图像会说话的人脸动画**

张文轩\*,1小东村\*,2 王宣3 张勇2 西申2 郭玉1 樱山2 王飞†,1

1西安交通大学 2腾讯人工智能实验室 3蚂蚁群

<https://sadtalker.github.io>

***输入音频***

***样式1***

***样式2***

***样式3***

图1.我们提出的SadTalker能够根据输入的音频和单张参考图片生成多样化、逼真且同步的说话视频。

**摘要**

*通过人脸图像和语音音频生成头部语音视频仍然包含许多挑战。即，不自然的头部运动，扭曲的表情，和身份的变化。我们认为，这些问题主要是由于从耦合的二维运动场中学习所致。另一方面，明确使用三维信息也会出现表达僵硬和视频不连贯的问题。我们提出了SadTalker，它从音频中生成3DMM的三维运动系数（头部姿态，表达），并为说话头部生成隐式调制一个新的三维感知面部渲染模型。*

\*相等贡献

†通讯作者

*为了学习真实的运动系数，我们明确地分别建模音频和不同类型的运动系数之间的联系。准确地说，我们提出了ExpNet，通过提取系数和3d渲染的面部表情来从音频中学习准确的面部表情。对于头部姿态，我们通过条件VAE设计了PoseVAE来合成不同风格的头部运动。最后，将生成的三维运动系数映射到提出的人脸渲染的无监督三维关键点空间，并合成最终的视频。我们进行了大量的实验来证明我们的方法在运动和视频质量方面的优越性。*

**1.介绍**

用语音音频制作静态肖像图像是一项具有挑战性的任务，在数字人类创作、视频会议等领域有着许多重要的应用。以往的工作主要集中在产生嘴唇运动[[2,](#bookmark2)[3,](#bookmark3)[30,](#bookmark4)[31,](#bookmark5)[51](#bookmark6)]因为它与语言有很强的联系。最近的工作还旨在生成一个真实的谈话面部视频，包含其他相关的动作，如头部姿势。他们的方法主要是通过标记来引入二维运动场[[52](#bookmark7)]和隐式形变[[39,](#bookmark8)[40](#bookmark9)]。然而，生成的视频的质量仍然是不自然的，并受到特定姿势的限制[[17,](#bookmark10)[51](#bookmark6)]，月度模糊[[30](#bookmark4)]、身份修改[[39,](#bookmark8)[40](#bookmark9)]，和面部扭曲[[39,](#bookmark8)[40,](#bookmark9)[49]](#bookmark11)。

由于音频和不同动作之间的关系不同，生成一个自然的头部视频包含许多挑战。也就是说，嘴唇的运动与音频有最强的联系，但音频可以通过不同的头部姿势和眼睛的眨眼表达出来。因此，以前的面部表情，基于标记的方法[[2，](#bookmark2)[52](#bookmark7)]和基于2D流的音频表达网络[[39,](#bookmark8)[40](#bookmark9)]可能会产生扭曲的面部，因为头部的运动和表达在其表征中没有完全解耦开。另一种流行的方法类型是基于潜在的人脸动画[[3,](#bookmark3)[17,](#bookmark10)[30,](#bookmark4)[51](#bookmark6)]他们的方法主要集中在说话人脸动画中的特定动作类型，难以合成高质量的视频。我们的观察是，三维面部模型包含一个高度解耦的表示，可以用来单独学习每种类型的运动。虽然相似的话题也被讨论过[[49](#bookmark11)]，但是他们的方法仍然产生了不准确的表达和不自然的运动序列。

从以上观察，我们提出了SadTalker，一个通过隐式

三维系数调制的音频驱动语音头视频生成系统(Stylized

Audio-Driven Talking-head video generation

system)。为了实现这一目标，为了实现这一目标，我们

将3DMM的运动系数作为中间表示，并将我们的任务划分为

两个主要组成部分。一方面，我们的目标是从音频生成现

实的运动系数（例如，头姿势，嘴唇运动，眨眼）并且分

别学习每个运动来减少不确定性。对于表情生成，我们设

计了一个新颖的从音频到表情系数的网络，通过从[30]的唇

部运动系数中提取系数，并在重构的渲染3D面部[5]上应用

感知损失（如唇读损失[1]、面部标记点损失）来实现。对

于具有风格化的头部姿势，我们使用条件变分自编码器

（VAE）[6]通过学习给定姿势的残差来模拟多样化和逼真

的头部运动。在生成逼真的3DMM系数后，我们通过一个新

颖的3D感知面部渲染器驱动源图像。受到face-vid2vid[42]

的启发，我们学习了明确的3DMM系数与无监督3D关键点域

之间的映射关系。然后，通过源图像和驱动图像的无监督

3D关键点生成变形场，它将参考图像变形以生成最终视频。我们分别对表情生成、头部姿势生成和面部渲染的各个子

网络进行训练，而我们的系统可以以端到端的方式进行推

断。关于实验，多个指标显示了我们方法在视频和动作质

量方面的优势。

本文的主要贡献可以总结为：

* 我们提出了SadTalker，一个使用生成的真实三维运动系数的新颖系统风格的音频驱动的单图像说话人脸动画模型。
* 为了从音频中学习3DMM模型的真实三维运动系数，本文分别给出了ExpNet和PoseVAE。
* 提出了一种新的语义解纠缠和三维感知的人脸渲染方法来制作真实的说话头部视频。
* 实验结果表明，该方法在运动同步和视频质量方面都取得了最先进的性能。

**2.相关工作**

**音频驱动的单个图像说话的人脸生成**。早期工作[[3,](#bookmark3)[30,](#bookmark4)[31](#bookmark5)]主要专注于用知觉鉴别器产生准确的嘴唇运动。由于真实的视频包含了许多不同的动作，ATVGnet [[2]](#bookmark2)使用面部标记作为中间表示法来生成视频帧。MakeItTalk也提出了类似的方法[[52](#bookmark7)]，不同的是，它将从输入声信号中分离出内容和扬声器信息。由于面部标记仍然是一个高度耦合的空间，在解纠缠的空间中产生会说话的头也是最近流行的。PC-AVS[[51](#bookmark6)]使用隐式潜在代码解耦头部姿态和表情。然而，它只能产生低分辨率的图像，并需要从另一个视频中获得控制信号。Audio2Head[39]和Wang等人[40]从视频驱动方法[36]中获得灵感，用以生成会说话的头部面部图像。然而，这些头部运动仍然不生动，并且会产生扭曲的面部，识别不准确。虽然有一些以前的作品[[33,](#bookmark17)[49]](#bookmark11)使用3DMM作为中间表示，他们的方法仍然面临着表情生成不准确的问题[[33](#bookmark17)]和明显的人工痕迹[[49]](#bookmark11).

**音频驱动的视频人像。**我们的任务还与视觉配音有关，其目的是通过音频编辑一个人像视频。与音频驱动的单张图片说话人脸生成不同，该任务通常需要在特定的视频上进行训练和编辑，遵循之前的深度视频肖像工作[[19](#bookmark18)]，这些方法利用3DMM信息进行人脸重建和动画制作[[45](#bookmark19)], NVP[38]和 AD-NeRF[11]学习再现表情来编辑嘴型。除了唇部运动，例如头部动作[23, 48]和情感性的说话面部[18]也受到关注。基于3DMM的方法在这些任务中扮演着重要角色，因为它可以实际地从视频片段中拟合3DMM参数。尽管这些方法在个性化视频中取得了令人满意的结果，但它们的方法不能应用于任意照片和自然环境中的音频



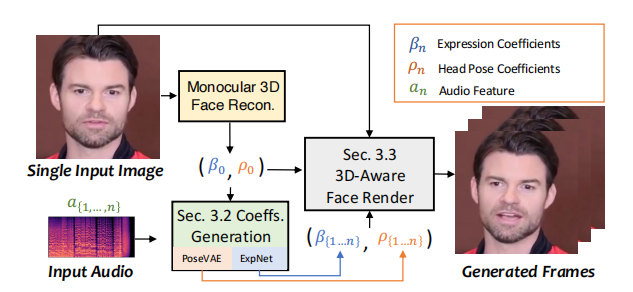


图2.我们的方法使用3DMM的系数作为中间运动表示。为此，我们首先从音频中生成真实的三维运动系数（面部表情β，头部姿态ρ），然后使用这些系数为最终的视频生成隐式调制三维感知的面部渲染模型。

**视频驱动的单个图像，会说话的人脸生成。**这项任务也被

称为人脸再现或人脸动画，其目的是将源图像的运动转移

到目标人物，它已经被广泛探索[[14,](#bookmark25)[29,](#bookmark26)[33,](#bookmark17)[36,](#bookmark16)[37,](#bookmark27)[41,](#bookmark28)[42,](#bookmark15)[44,](#bookmark29)[47,](#bookmark30)[50](#bookmark31)]最近，以前的作品也从源图像和目标中学习了一个共享的中间运

动表示，大致可以分为标记[[41](#bookmark28)]和无监督的标记方法[[36](#bookmark16)[,42,](#bookmark15)[50](#bookmark31)]，3DMM的方法[[7,](#bookmark32)[33,](#bookmark17)[47](#bookmark30)]和潜在的动画[[25,](#bookmark33)44][。这项任务比我们的任务要容易得多](#bookmark29)，因为它包含在同一域中的运动。我们的面

部渲染方法同样受到了基于无监督特征点方法[42]和基于

3DMM方法[33]的启发，用于映射学习到的系数以生成真实视

频。然而，它们并不专注于生成逼真的运动系数。

**3.方法**

如图2所示，我们的系统使用3D运动系数作为生成说话头部的中间表示。我们首先从原始图像中提取这些系数。接着，ExpNet和PoseVAE分别生成逼真的3DMM运动系数。最后，提出了一个3D感知的面部渲染器来生成说话头部视频。在第3.1节，我们简要介绍了作为预备知识的3D面部模型，以及在第3.2节和第3.3节分别设计的音频驱动运动系数生成和系数驱动的图像动画器。

**3.1.三维人脸模型的初步研究**

由于三维信息是在三维环境中捕获的，因此它对于提高生成视频的真实性是至关重要的。然而，以前的工作[[30,](#bookmark4)[51,](#bookmark6)[52](#bookmark7)]在三维空间中，由于很难从单一图像中获得精确的三维系数，而且高质量的人脸渲染也难以设计，因此很少被考虑。受最近的单图像深度三维重建方法的启发[[5](#bookmark13)]，我们将预测的三维变形模型（3DMM）的空间作为我们的中间表示。在3DMM中，三维面形状S可以解耦为：.

 (1)

其中S是3D面部的平均形状，Uid和Uexp是LSFM可变形模型[1]的身份和表情的正交基。系数和分别描述了个人身份和表情。为了保留姿势的变化，系数和表示头部的旋转和平移。为了实现与身份无关的系数生成[33]，我们仅模拟运动的参数。我们从驱动音频中分别学习头部姿势和表情系数β，如前所述。然后，这些运动系数被用来隐式调制我们的面部渲染器，以生成最终视频。

**3.2.通过音频产生的运动系数**

如上所述，三维运动系数同时包含头部姿态和表达式，其中头部姿态为全局运动，表达式为相对局部运动。为此，完全学习一切都会导致巨大的网络类型，因为头部的姿势与音频的关系相对较弱，而嘴唇的运动是高度连接的。我们使用下面分别介绍的PoseVAE和ExpNet来生成头部姿态和表达式的运动。

**ExpNet** 训练一个从音频中产生精确的表达式系数的通用模型是非常困难的，原因有两个：1)音频到表情并不是一个针对不同身份的一对一映射任务。2)在表情系数中存在一些与听觉无关的运动，它会影响预测的准确性。我们的计算方法旨在减少这些不确定性。至于身份问题，我们通过第一帧的表达系数β将表达运动与特定的人联系起来。为了在自然对话中减少其他面部组件的运动权重，我们通过预训练的Wav2Lip[30]网络和深度3D重建[5]技术，使用仅限唇部运动的系数作为目标系数。然后，可以通过对渲染图像上的附加标记点损失来利用其他较小的面部动作（例如，眨眼）。

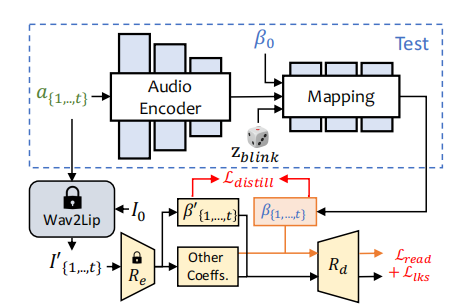


图3。我们的扩展网络的结构。我们采用了一个单眼三维人脸重建模型[[5](#bookmark13)](Re和Rd)来学习实际的表达式系数。其中Re是一个预先训练过的3DMM系数估计器，Rd是一个没有可学习参数的可微三维面渲染，我们使用参考表达式减少身份的不确定性并且从预先训练好的Wav2Lip中生成的帧[[30](#bookmark4)]和第一帧作为目标表达式系数，因为它只包含与嘴唇相关的运动。

如图3所示，我们从音频窗口生成t帧表情系数，

其中每帧的音频特征是一个0.2秒的梅尔频谱图。在训练过

程中，我们首先设计了一个基于ResNet的音频编码器

[12, 30]，将音频特征嵌入到一个潜在空间中。然后，添加一个线性层作为映射网络来解码表情系数。在这里，

我们还添加了从参考图像中获得的参考表情来减少之前

讨论过的身份不确定性。由于我们在训练中使用仅唇部的

系数作为真实数据，我们显式地添加了一个眨眼控制信号和相应的眼部标记点损失，以生成可控的眨

眼动作。形式上，网络可以写成：

 (2)

在损失函数方面，我们首先使用来评估仅唇部表情系与生成的之间的差异。请注意，我们只使用wav2lip的第一帧来生成唇部同步视频，这减少了姿势变化和除唇部运动以外的其他面部表情的影响。此外，我们还涉及可微分的3D面部渲染来计算明确的面部运动空间中的额外感知损失。如图3所示，我们计算标记点损失来衡量眨眼范围和整体表情准确性。一个预训练的唇读网络也被用作时序唇读损失，以保持感知唇部质量[9, 30]。我们在补充材料中提供了更多训练细节。

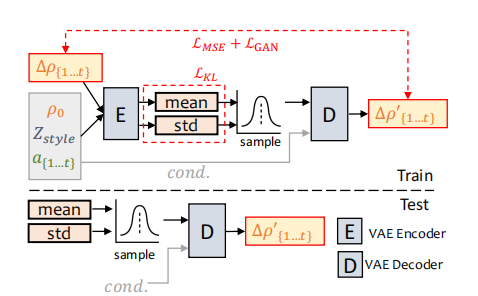


图4。我们提出的PoseVAE处理流程。我们通过一个条件VAE结构学习输入头部姿势的残差。给定条件：第一帧姿势、风格身份和音频片段，我们的方法学习头部姿势残差的分布。训练后，我们可以仅通过姿势解码器和条件(cond.)生成风格化的结果。

**PoseVAE** 如图4所示，我们设计了一个基于VAE[21]的模型，用于学习真实对话视频中逼真且具有身份特征的风格化头部运动。在训练过程中，姿势VAE在固定的n帧上使用基于编码器-解码器的结构进行训练。编码器和解码器都是双层的多层感知机（MLP），其中输入包含一个序列化的t帧头部姿势，我们将其嵌入到一个高斯分布中。在解码器中，网络学习从采样分布生成t帧姿势。我们的PoseVAE不是直接生成姿势，而是学习第一帧条件姿势的残差，这使得我们的方法在测试中能够在第一帧的条件下生成更长、更稳定、更连贯的头部运动。此外，根据CVAE[6]，我们添加了相应的音频特征和风格身份作为条件，以实现节奏感知和身份风格。KL散度用于衡量生成运动的分布。均方误差损失和对抗性损失用于确保生成质量。我们在补充材料中提供了更多关于损失函数的细节。

**3.3.3D感知人脸渲染**

在生成逼真的3D运动系数之后，我们通过一个精心设计的3D感知图像动画器渲染最终视频。我们从最近的图像动画方法face-vid2vid[42]中获得灵感，因为它能隐式地从单张图像中学习3D信息。然而，他们的方法需要一个真实视频作为运动驱动信号。我们的面部渲染器使其可以通过3DMM系数来驱动。

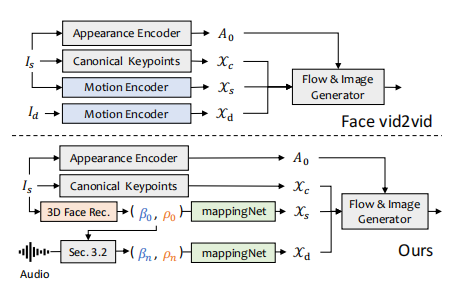


图5。我们提出的FaceRender以及与face-vid2vid[42]的比较。给

定源图像和驱动图像，face-vid2vid在无监督的3D关键点

空间、和中生成运动。然后，可以通过外观和关键点生成图像。由于我们没有驱动图像，我们使用明确分离的3DMM

系数作为代理，并将其映射到无监督的3D关键点空间。

如图5所示，我们提出了mappingNet来学习明确的3DMM运动系数（头部姿势和表情）与隐式的无监督3D关键点之间

的关系。我们的mappingNet是通过几个1D卷积层构建的。

我们使用来自时间窗口的临时系数进行平滑处理，如PIR-

enderer[33]所示。不同的是，我们发现在PIRenderer中的面部对齐运动系数会极大地影响音频驱动视频生成的运动自然性，并在第4.4节提供了一个实验。我们只使用表情和头部姿势的系数。

在训练方面，我们的方法包含两个步骤。首先，我们按照原始论文中的方法，以自监督的方式训练face-vid2vid[42]。在第二步，我们冻结外观编码器、规范关键点估计器和图像生成器的所有参数，用于调优。然后，我们在重建风格中，针对真实视频的3DMM系数训练映射网络。我们使用L1损失在无监督关键点的领域内进行监督，并按照他们原始的实现方式生成最终视频。更多细节可以在补充材料中找到。

**4.实验**

**4.1.实施细节和指标**

**Datasets** 我们使用VoxCeleb[[26](#bookmark44)]训练数据集，包含1251名

受试者的超过10w个视频。我们按照以前的图像动画方法

裁剪原始视频[[36](#bookmark16)]，并将视频的大小调整为256×256。经过

预处理后，这些数据被用于训练我们的face渲染器。由于

一些视频和音频在VoxCeleb中没有对齐，我们选择了

1890个对齐的视频和音频来训练我们的PoseVAE和ExpNet。输入的音频被降采样到16 kHz，并转换为与 Wav2lip相同设置的梅尔光谱图[[30](#bookmark4)]。为了测试我们的方法，我们使用了来自HDTF数据集的346个视频的第一个8秒视频（总共约7万帧）[[49](#bookmark11)]因为它包含了高分辨率谈话头部视频。这些视频也按照[36]的方法进行了裁剪和处理，并且为了评估目的被调整大小至256×256像素。我们使用每个视频的第一帧作为参考图像来生成视频。

**实现细节** ExpNet、PoseVAE和FaceRender都是分别训练的，我们在所有实验中都使用了Adam优化器[20]。经过训练，我们的方法可以以端到端的方式推断，无需任何人工干预。所有的3DMM（三维可变形模型）参数都是通过预训练的深度3D人脸重建方法[5]提取的。我们在8块A100GPU上进行所有实验。ExpNet、PoseVAE 和FaceRender的学习率分别为、和。就时间因素考虑，ExpNet 使用连续的5帧来学习。PoseVAE 通过连续的32帧来学习。FaceRender 中的帧是逐帧生成的，使用连续5帧的系数来保持稳定性。

**评价指标** 我们通过多个在以往研究中广泛使用的指标来展示我们方法的优越性。我们采用弗雷歇特初始距离（FID）[13, 35]和累积概率模糊检测（CPBD）[27]来评估图像的质量，其中FID用于评估生成帧的真实感，而CPBD用于评估生成帧的清晰度。为了评估身份保持，我们计算了源图像与生成帧之间身份嵌入的余弦相似性（CSIM），在此过程中，我们使用ArcFace[4]来提取图像的身份嵌入。为了评估唇部同步和嘴形，我们从Wav2Lip[30]评估嘴形的感知差异，包括距离分数（LSE-D）和置信度分数（LSE-C）。我们还进行了一些度量来评估生成帧的头部运动。对于生成的头部运动的多样性，我们计算了使用Hopenet[28]从生成帧中提取的头部运动特征嵌入的标准偏差。对于音频和生成的头部运动的对齐，我们计算了如Bailando[22]中的节奏对齐分数。

**4.2.与其他最先进的方法进行比较**

我我们比较了几种最先进的方法，用于生成说话头像视频（MakeItTalk[52]、Audio2Head[39]以及Wang等人的方法[40]）和音频到表情生成（Wav2Lip[30]、PC-AVS[51]）。评估是直接在它们公开可用的检查点上进行的。

1这种方法需要从音频中提取音素，这只对特定语言的有效。

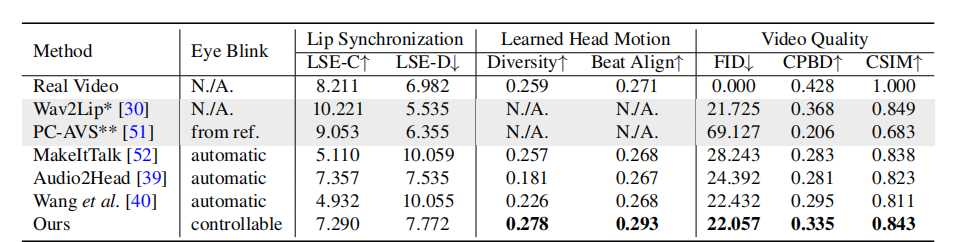


表1。在HDTF数据集上与最先进方法的比较。我们在单次设置中评估了Wav2Lip[30]和PC-AVS[51]。Wav2Lip\* 在视频质量上表现最佳，因为它只对唇部区域进行动画处理，而其他区域与原始帧相同。在使用固定参考姿势评估时，PC-AVS\*\*在一些样本中失败。



图6。我们将我们的方法与几种最先进的单图像音频驱动说话头部生成方法进行了比较。在嘴唇同步、身份保留、头部运动和图像质量方面，我们的方法产生了更高质量的结果。我们在上方提供了目标图像，用于嘴唇形状和身份参考。请参考我们的补充视频以进行更好的比较。

如表1所示，提出的方法在视频质量和头部姿势多样性方面表现更好，并且在唇部同步指标方面与其他完全的说话头像生成方法表现相当。我们认为这些唇部同步的指标对音频过于敏感，其中不自然的唇部运动可能会得到更高的分数。

然而，我们的方法与真实视频获得了相似的分数，这展示了我们的优势。我们还在图6中展示了不同方法的视觉结果。在这里，我们提供了唇部参考，以可视化我们方法的唇部同步效果。从图中可以看出，我们的方法在视觉质量上与原始目标视频非常相似





我们的

参考图像

联合学习姿态和经验系数

图7.我们将我们的方法与一个基准方法进行比较，该方法从一个不包含任何条件的单一网络中学习所有系数（来自 Speech2Gesture[10]）。我们的方法展现了清晰的头部运动、身份保持和多样的表情。

并且具有我们预期的不同头部姿势。与其他方法相比，Wav2Lip[30]生成了模糊的半面脸。PC-AVS[51]和 Audio2Head[39]在身份保持方面存在困难。Audio2Head 只能生成正面的说话脸部。此外，MakeItTalk[52]和 Audio2Head[39]由于2D变形而生成了扭曲的面部视频。我们在补充资料中提供了视频比较，以展示更清晰的对比。

**4.3.用户研究**

我们进行了用户研究，以评估所有方法的性能。我们生成了总共20个视频作为测试。这些样本包含几乎相等的性别，以及不同的年龄、姿势和表情，以展示我们方法的稳健性。我们邀请了20名参与者，让他们根据视频清晰度、唇部同步、头部运动的多样性和自然性以及整体质量来选择最佳方法。结果显示在表2中，参与者主要因为视频和运动质量而喜欢我们的方法。我们还发现，38%的参与者认为我们的方法在唇部同步方面比其他方法更好，这与表1的结果不一致。我们认为这可能是因为大多数参与者关注视频的整体质量，其中模糊和静止的面部视频[30, 51]影响了他们的意见。

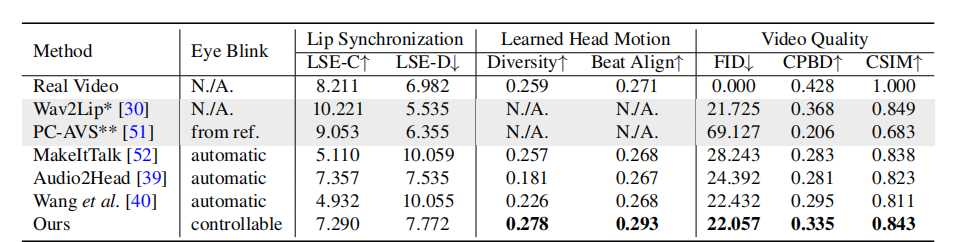


表2.用户研究。

**4.4.消融研究**

**ExpNet消融** 对于ExpNet，我们主要通过唇部同步指标来评估每个组件的必要性。

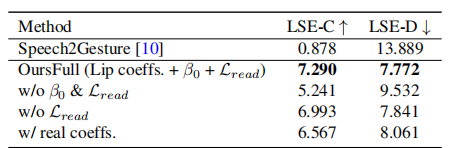


表3.对ExpNet的消融实验。初始表情和唇读损失都极大地提高了性能。然而，在使用真实系数时，唇部同步指标大幅下降。



图8.对ExpNet的消融实验。我们从生成的视频中选择四帧进行比较。我们的方法大大减少了从音频到表情生成的不确定性。参考用于提供身份信息，而仅唇部的系数则生成了更好的唇部同步效果。请注意，目标图像作为身份和唇部运动的参考提供。

由于之前没有解耦的方法，我们考虑了一个基线方法（Speech2Gesture[10]，这是一个从音频到关键点生成的网络）来共同学习头部姿势和表情系数。如表3和图7所示，学习所有的运动系数是很难生成可信的说话头像视频的。然后，我们考虑了提出的ExpNet的变体，初始表情、唇读损失和仅唇部系数的必要性都是关键。视觉比较显示在图8中，我们的方法在没有初始表情的情况下显示出预期的巨大身份变化。此外，如果我们使用真实系数来替换我们使用的仅唇部系数，唇部同步的性能会大幅下降。

**PoseVAE消融** 我们从运动多样性和音频节奏对齐的角度评估了提出的PoseVAE。如表4所示，基线方法Speech2Gesture[10]在姿势评估方面的表现也较差。

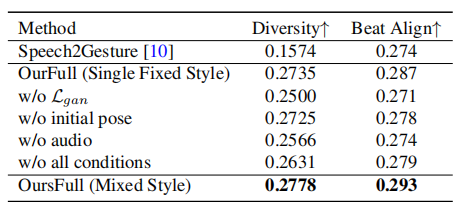


表4.对提出的PoseVAE的多样性和音频对齐进行消融研究。每个组件或条件都极大地有助于生成真实的头部运动。

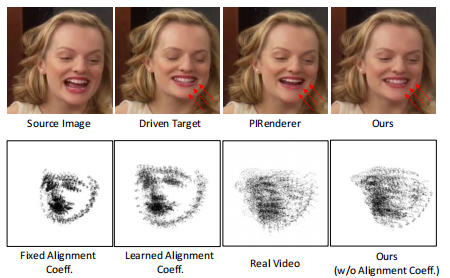


图9.面部渲染的消融研究。在第一行，我们直接将我们的方法与PIRenderer[33]进行面部动画的比较，我们的方法显示了更好的表情建模。第二行是从相同运动系数生成的面部标记的轨迹图。使用额外的面部对齐系数作为运动系数的一部分[33]会生成不真实的对齐头部视频。

对于我们的变体，由于我们的方法包含了几种身份风格标签，为了更好地评估其他组件，我们首先考虑在我们完整方法的一个固定单热风格下进行消融研究以进行评估（我们的完整方法，单一固定风格）。我们设置中的每个条件都有助于提高整体运动质量，无论是在多样性还是节奏对齐方面。我们进一步报告了我们完整方法的混合风格的结果，它使用随机选择的身份标签作为风格，并且在多样性表现上也表现更好。由于姿势差异在图中难以显示，请参阅我们的补充材料以进行更好的比较。

**面部渲染消融** 我们从两个方面对提出的面部渲染进行了消融研究。一方面，我们展示了我们的方法与PIRenderer[33]的重建质量，因为这两种方法都使用3DMM作为中间表示。如图9的第一行所示，提出的面部渲染由于稀疏无监督关键点的映射，展示了更好的表情重建质量。准确的表情映射也是实现唇部同步的关键。



图10.限制。在某些示例中，我们的方法可能会在唇部区域显示一些牙齿伪影，可以通过面部修复网络来改进，即GFPGAN [43]。

此外，我们评估了PIRenderer[33]中使用的额外对齐系数引起的姿势不自然。如图9的第二行所示，我们绘制了使用相同头部姿势和表情系数生成的视频中标记的轨迹图。使用固定或可学习的裁剪系数（作为我们的PoseVAE中姿势系数的一部分）会生成面部对齐的视频，这在自然视频中看起来很奇怪。我们移除它并直接使用头部姿势和表情作为调制参数，显示出更真实的结果。

**4.5.限制**

尽管我们的方法可以从单张图片和音频中生成真实的视频，但我们的系统仍然存在一些局限性。由于3DMMs（三维形态建模）不模拟眼睛和牙齿的变化，我们的面部渲染中的mappingNet在某些情况下也难以合成真实的牙齿。这个局限可以通过盲目面部恢复网络[43]来改善，如图10所示。我们工作的另一个局限是，我们只关注唇部运动和眼睛眨眼，而不是其他面部表情，例如情绪和凝视方向。因此，生成的视频具有固定的情绪，这也降低了生成内容的真实性。我们将其视为未来的工作方向。

**5.结论**

在这篇论文中，我们提出了一个新的风格化音频驱动的说话头像视频生成系统。我们使用3DMM（三维形态模型）的运动系数作为中间表示，并学习它们之间的关系。为了从音频生成真实的3D系数，我们提出了ExpNet和PoseVAE，用于生成真实的表情和多样的头部姿势。为了模拟3DMM运动系数与真实视频之间的关系，我们提出了一个新颖的3D感知面部渲染方法，灵感来源于图像动画方法[42]。实验展示了我们整个框架的优越性。由于我们预测了真实的3D面部系数，我们的方法也可以直接用于其他模式，即个性化的2D视觉配音[45]、2D卡通动画[52]、3D面部动画[8,46]以及基于NeRF的4D说话头像生成[15]。

**伦理考虑** 我们考虑了提出的方法的误用，因为它可以从单张面部图像生成非常真实的视频。类似于Dall-E [32]和Imagen[34]，可见和不可见的视频水印将被插入到生成的视频中，用于生成内容的识别。我们也希望我们的方法能在伪造检测领域提供新的研究样本。

**参考文献**

[1] Volker Blanz and Thomas Vetter. A morphable model for the

synthesis of 3d faces. In ACM SIGGRAPH, 1999. 2, 3

[2] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang

Xu. Hierarchical cross-modal talking face generation with

dynamic pixel-wise loss. In CVPR, 2019. 2

[3] Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei

Yin, Mingrui Zhu, Xuan Wang, Jue Wang, and Nannan Wang.

Videoretalking: Audio-based lip synchronization for talking

head video editing in the wild. In SIGGRAPH Asia 2022

Conference Papers, pages 1–9, 2022. 2

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou.

Arcface: Additive angular margin loss for deep face recogni

tion. In CVPR, 2019. 5

[5] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde

Jia, and Xin Tong. Accurate 3d face reconstruction with

weakly-supervised learning: From single image to image set.

In CVPR Workshops, 2019. 2, 3, 4, 5, 12, 13, 14

[6] Carl Doersch. Tutorial on variational autoencoders. arXiv

preprint arXiv:1606.05908, 2016. 2, 4

[7] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia

Sharmanska. Headgan: One-shot neural head synthesis and

editing. In ICCV, 2021. 3

[8] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and

Taku Komura. Faceformer: Speech-driven 3d facial animation

with transformers. In CVPR, 2022. 8

[9] Panagiotis P. Filntisis, George Retsinas, Foivos Paraperas

Papantoniou, Athanasios Katsamanis, Anastasios Roussos,

and Petros Maragos. Visual speech-aware perceptual 3d fa

cial expression reconstruction from videos. arXiv preprint

arXiv:2207.11094, 2022. 4, 13

[10] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, An

drew Owens, and Jitendra Malik. Learning individual styles

of conversational gesture. In CVPR, 2019. 7, 8, 14

[11] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun

Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radi

ance fifields for talking head synthesis. In ICCV, 2021. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.

Deep residual learning for image recognition. In CVPR, 2016. 4

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner,

Bern hard Nessler, and Sepp Hochreiter. Gans trained by a two

time-scale update rule converge to a local nash equilibrium.

In NeurIPS, 2017. 5

1. Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu.

Depth aware generative adversarial network for talking head

video generation. In CVPR, 2022. 3

1. Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and

Juyong Zhang. Headnerf: A real-time nerf-based parametric

head model. In CVPR, 2022. 8

1. Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A

Efros. Image-to-image translation with conditional adversarial

net works. CVPR, 2017. 14

1. Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne

Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional

talking face via audio-based emotion-aware motion model. In

ACM SIGGRAPH, 2022. 2

[18] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu,

Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven

emotional video portraits. In CVPR, 2021. 3

[19] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng

Xu, Justus Thies, Matthias Niessner, Patrick Perez, Christian

Richardt, Michael Zollhofer, and Christian Theobalt. Deep

video portraits. ACM Transactions on Graphics (TOG), 2018. 2

[20] Diederik P Kingma and Jimmy Ba. A method for stochastic

optimization. arXiv preprint arXiv:1412.6980, 2014. 5

[21] Diederik P Kingma and Max Welling. Auto-encoding varia

tional bayes. CoRR, abs/1312.6114, 2014. 4

[22] Siyao Li, Yu Weijiang, Gu Tianpei, Lin Chunze, Wang Quan,

Qian Chen, Loy Chen Change, and Liu Ziwei. Bailando 3d dance

generation via actor-critic gpt with choreographic memory. In

CVPR, 2022. 5

[23] Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live speech por

traits: real-time photorealistic talking-head animation. ACM

Transactions on Graphics (TOG), 2021. 3

[24] Pingchuan Ma, Yujiang Wang, Stavros Petridis, Jie Shen, and

Maja Pantic. Training strategies for improved lip-reading. In

ICASSP, 2022. 13

[25] Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Implicit

Warping for Animation with Image Sets. In NeurIPS, 2022. 3

[26] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman.

Voxceleb: a large-scale speaker identifification dataset. In

INTERSPEECH, 2017. 5, 11, 12

[27] Niranjan D. Narvekar and Lina J. Karam. A no-reference

image blur metric based on the cumulative probability of blur

detection (cpbd). TIP, 2011. 5

[28] Ruiz Nataniel, Eunji Chong, and Rehg James M. Fine-grained

head pose estimation without keypoints. In CVPR Workshops,

2018. 5

[29] Youxin Pang, Yong Zhang, Weize Quan, Yanbo Fan, Xi

aodong Cun, Ying Shan, and Dong-ming Yan. Dpe: Disen

tanglement of pose and expression for general video portrait

editing. arXiv preprint arXiv:2301.06281, 2023. 3

[30] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P.Namboodiri,

and C.V.Jawahar. A lip sync expert is all you need for speech

to lip generation in the wild. In ACM MM, 2020. 2, 3, 4, 5, 6,

7, 11, 12

[31] Prajwal K R, Rudrabha Mukhopadhyay, Jerin Philip, Ab

hishek Jha, Vinay Namboodiri, and C V Jawahar. Towards

automatic face-to-face translation. In ACM MM, 2019. 2

1. Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey

Chu, and Mark Chen. Hierarchical text-conditional image

generation with clip latents. arXiv preprint arXiv:2204.06125,

2022. 9

[33] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan

Liu. Pirenderer: Controllable portrait image generation via

semantic neural rendering. In ICCV, 2021. 2, 3, 5, 8, 11, 14

[34] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour,

Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes,

et al. Photorealistic text-to-image diffusion models with deep

language understanding. arXiv preprint arXiv:2205.11487,

2022. 9

[35] Maximilian Seitzer. pytorch-fifid: FID Score for PyTorch.

https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.2.1. 5

1. Aliaksandr Siarohin, Stephane Lathuiliere, Sergey

Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model

for image animation. In NeurIPS, 2019. 2, 3, 5, 11

[37] Aliaksandr Siarohin, Oliver Woodford, Jian Ren, Menglei

Chai, and Sergey Tulyakov. Motion representations for artic

ulated animation. In CVPR, 2021. 3

[38] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian

Theobalt, and Matthias Nießner. Neural voice puppetry:

Audio-driven facial reenactment. In ECCV, 2020. 2

[39] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and

Xin Yu. Audio2head: Audio-driven one-shot talking-head

generation with natural head motion. In IJCAI, 2021. 2, 5, 6,

7, 12

[40] Suzhen Wang, Lincheng Li, Yu Ding, and Xin Yu. One

shot talking face generation from single-speaker audio-visual

correlation learning. In AAAI, 2022. 2, 5, 6, 7, 12

1. Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan

Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis.

In NeurIPS, 2019. 3

[42] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot

free-view neural talking-head synthesis for video conferenc

ing. In CVPR, 2021. 2, 3, 4, 5, 8, 11, 14

[43] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. To

wards real-world blind face restoration with generative facial

prior. In CVPR, 2021. 8

[44] Yaohui Wang, Di Yang, Francois Bremond, and Antitza

Dantcheva. Latent image animator: Learning to ani

mate images via latent space navigation. arXiv preprint

arXiv:2203.09043, 2022. 3

[45] Xin Wen, Miao Wang, Christian Richardt, Ze-Yin Chen, and

Shi-Min Hu. Photorealistic audio-driven video portraits. IEEE

Transactions on Visualization and Computer Graphics, 26(12):3457–

3466, 2020. 2, 8

[46] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun,

Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven

3d facial animation with discrete motion prior. arXiv preprint

arXiv:2301.02379, 2023. 8

[47] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo

Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and

Yujiu Yang. Styleheat: One-shot high-resolution editable

talking face generation via pre-trained stylegan. In ECCV,

2022. 3

[48] Chenxu Zhang, Yifan Zhao, Yifei Huang, Ming Zeng,

Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo. Facial:

Synthesizing dynamic talking face with implicit attribute

learning. In ICCV, 2021. 3

[49] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan.

Flow-guided one-shot talking face generation with a high

resolution audio-visual dataset. In CVPR, 2021. 2, 5, 11,

12

[50] Jian Zhao and Hui Zhang. Thin-plate spline motion model

for image animation. In CVPR, 2022. 3

[51] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy,

Xiaogang Wang, and Ziwei Liu. Pose-controllable talking

face generation by implicitly modularized audio-visual repre

sentation. In CVPR, 2021. 2, 3, 5, 6, 7, 12

[52] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevar

ria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk:

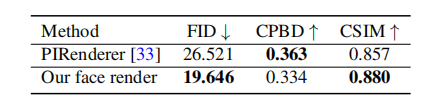
speaker-aware talking-head animation. ACM Transactions on

Graphics (TOG), 2020. 2, 3, 5, 6, 7, 8, 1

**A.附加实验**

**A.1.PIRenderer 与我们的面部渲染方法在面部再现方面的对比。**

我们比较了我们的面部渲染方法和PIRenderer[33] 在视频驱动的面部再现任务上的表现。我们已经在主论文的图9中展示了视觉对比，这里，我们给出了在HDTF数据集上的数值比较结果。我们使用跨身份设置评估这两种方法，并在354个视频上进行了结果分析。如表A5所示，提出的方法在FID和CSIM方面显示出更好的视觉质量，这证明了提出方法在音频驱动的说话头像生成方面的优势。对于这项任务中提出的方法和PIRenderer之间的更多差异，我们还在第B.4节讨论了面部对齐系数的影响。



表A5。面渲染评估。

**A.2.交叉ID设置和更多的测试数据集**

在主论文中，我们进行了相同身份的实验，其中测试视频的第一帧被视为参考图像，相应的音频被视为驱动信号，生成一个表情同步但头部姿势多样的视频。不同的是，在跨身份实验中，驱动音频来自另一个视频。这种设置也广泛用于视频驱动的面部再现的比较[36]。在跨身份实验中，PC-AVS的参考姿势来自音频对应的视频。

为此，除了主论文中的HDTF[49]数据集（我们在相同身份实验下评估结果），我们还在HDTF和VoxCeleb2[26] 数据集上以跨身份设置进行评估，如表A6和表A7所示。VoxCeleb2包含6112位发言者的超过100万次发言，其中测试集包含118位发言者的36k次发言。我们为每位发言者随机选择3个视频，获得354个视频进行评估。评估指标与HDTF数据集上的相同身份实验相同。我们直接在这个数据集上评估所有模型的预训练模型。我们以[36]中相同的方式裁剪视频，并将帧大小调整为256×256。如表中所示，提出的方法在这种设置下在大多数指标上在两个数据集上都显示出更好的唇部同步性。在最终视频的头部运动和视觉质量方面也观察到了相同的趋势。

**B.更多的实现细节**

我们在第B.1节、第B.2节、第B.3节和第B.4节提供了详细的音频预处理、网络结构、损失函数，以及对齐系数的讨论。

**B.1.音频预处理的细节**

我们遵循Wav2Lip[30]对音频进行预处理。具体来说，我们将所有音频预处理为16kHz。然后，我们将其转换为梅尔频谱图，使用FFT窗口大小为800，跳跃长度为200，以及80个梅尔滤波器组。因此，对于每一帧，我们有0.2秒的梅尔频谱图特征，其形状为16×80。

**B.2.网络结构详细信息**

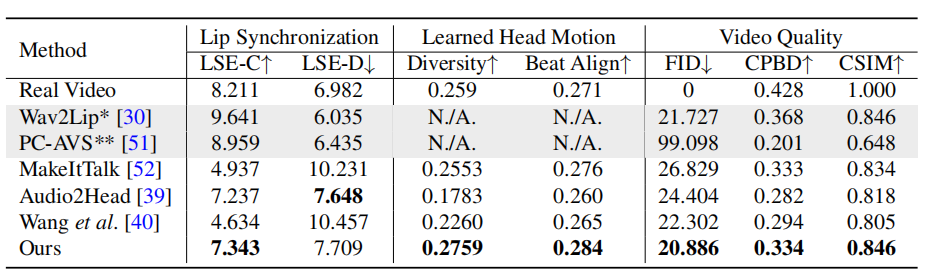
**ExpNet** 我们的ExpNet是通过一个音频编码器和一个线性层构建的。我们使用预训练的Wav2Lip的参数来初始化音频编码器。正如主论文中讨论的那样，我们首先将音频特征编码成音频嵌入。然后，我们生成表情系数，附加上第一帧的系数和眨眼控制信号。如图B11(c)所示，音频编码器是通过四个阶段的ResBlock-C构建的，如图B11(a)所示。我们只使用一个单一的线性层，如图B11(b)所示。

**PoseVAE** 如图B11(e)所示，我们的PoseVAE的编码器和解码器都包含几个线性层。对于条件，编码器和是通过连接、46维的独热向量（我们的训练数据集包含46个身份）和来自音频编码器的编码特征映射的。至于解码器，我们首先添加重新参数化的特征和样式嵌入。然后，我们连接音频特征，类似于编码器。

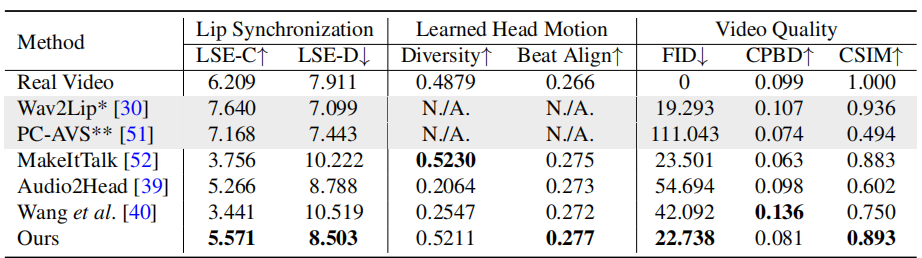
**FaceRender** 正如主论文的图5中讨论的那样，我们的面部渲染受到了运动转移方法face-vid2vid[42]的启发。我们引入了一个mappingNet来重新映射学习到的3DMM运动系数到无监督的3D关键点的空间。如图B11(d)所示，映射网络包含了t帧姿势和表情的运动系数，以在第t帧生成face-vid2vid[42]的运动表示（偏航、俯仰、滚动、平移和δ）。我们的FaceRender中的其他网络在[42]中有相同的结构。请参阅face-vid2vid[42]以获取更多有关FaceRender的网络细节。

**B.3.损失函数详细信息**

正如主论文中所述，我们使用从预训练的wav2lip[30]生成的表情系数



表A6。与HDTF数据集上最先进方法的比较[[49](#bookmark11)*带有跨身份设置。*Wav2Lip\*获得了最好的视频质量，因为它只动画的嘴唇区域，而其他区域是相同的原始帧。在交叉身份设置中，PC-AVS\*\*使用驾驶视频中的参考姿态进行评估，在一些样本中失败。



表A7。与VoxCeleb2的最新方法进行比较[[26](#bookmark44)*]在交叉身份设置下的数据集。*Wav2Lip\*获得了最好的视频质量，因为它只动画的嘴唇区域，而其他区域是相同的原始帧。在交叉身份设置中，PC-AVS\*\*使用驾驶视频中的参考姿态进行评估，在一些样本中失败。

然后执行3D面部捕捉[5]作为指导（简称唇部表情系数）。基本上，我们通过训练中唇部表情系数与生成的表情系数之间的均方误差来计算。形式上，一个T帧的可以写为：

 (3)

其中和分别是唇部专用和生成的表情系数。

我们还计算了渲染的3D面部投影的2D标记点上的损失函数

具体来说，如图B12所示，在第t帧中，眼睛区域的高度和宽度定义如下：

 (4)

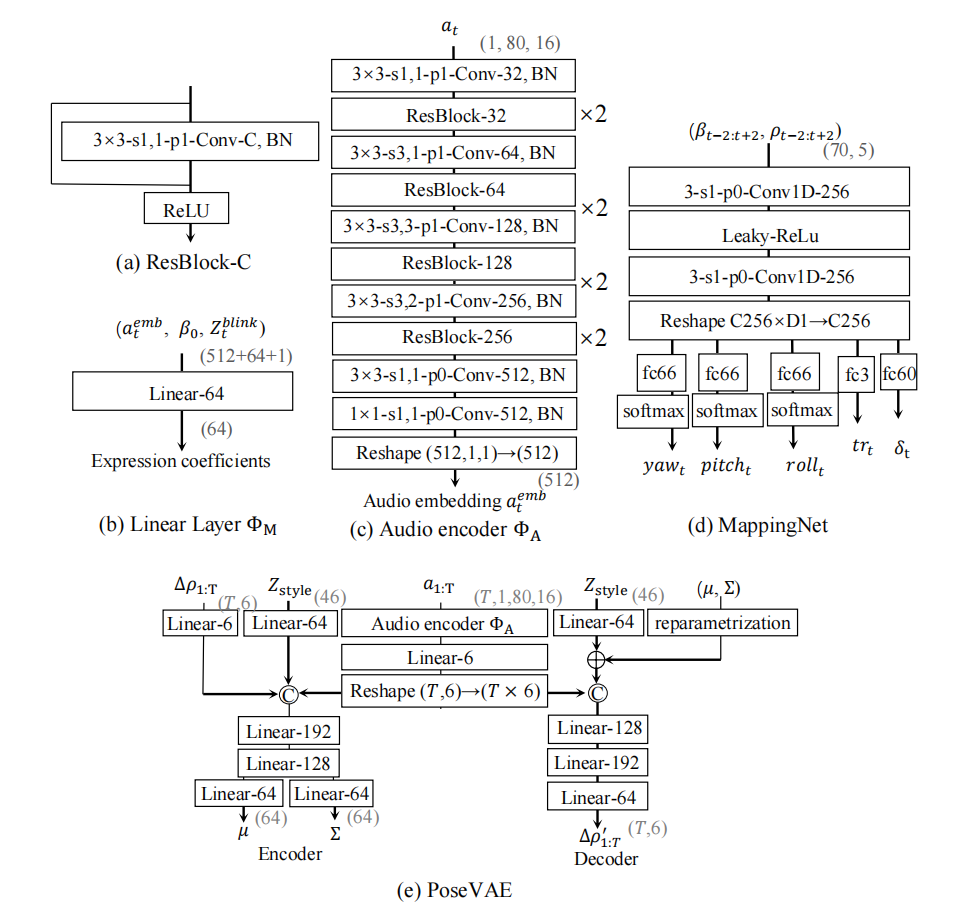
 (5)

 (6)

其中，是第t帧中眼睛区域的宽度，是第t帧中眼睛区域的高度，是第t帧中的第i个标记点。我们定义作为预测值，并按照以下方式计算眼睛损失：

 (7)

其中是第t帧的眼睛眨眼控制信号，它是均匀和随机生成的。为了消除对其他面部表情的影响，我们还限制了其他标记点的最小修改。



图B11。我们模型中网络的架构。这里，“3×3-s1,1-p1-Conv-32”表示一个卷积层，其核大小为3×3，步长大小为(1,1)，填充大小为(1,1)，输出通道为32。

因此:

 (8)

其中睛被设置为200，是由唇部专用表情系数预测的标记点，M是除眼睛区域外的一组标记点。

此外，此外，我们使用在[24]中提出的预训练唇读模型来计算由[9]启发的唇读损失。我们使用预训练的基于视频的唇读模型，其中输入是一系列（在我们的案例中为5帧）裁剪的嘴巴周围有趣区域的图像（如图B13所示），目标是一系列字符序列。因此，我们在ExpNet中使用一个可微分的3D面部渲染器[5]来通过生成的表情系数渲染图像，然后，我们使用嘴巴标记点的边界框裁剪渲染图像的嘴巴区域，获得字符序列的对数Cp。



‘”

‘”



‘"





‘"

图B12。面部标记可视化。



图B13。在三维渲染的人脸序列中裁剪的有趣区域的例子，以计算唇读损失。

至于监督，我们使用音频驱动的唇读模型从真实音频中生成字符序列的对数。因此，我们的目标是最小化和之间的差异。换句话说，

阅 () ()

(9)

总的来说，ExpNet的最终损失为：

 (10)

其中，lks,,, ，分别设置为2、0.01和0.01。

**PoseVAE** 我们首先计算重建损失在生成的之间计算均方损失和原来的{1...T }:

 (11)

同时，我们鼓励潜在空间分布与均值向量µ和协方差矩阵P的高斯分布的相似性。因此，我们将LKL定义为潜在空间分布与高斯分布之间的库尔贝克-莱布勒（KL）散度。我们还采用了基于PatchGAN[16]的判别器D，对头部运动序列执行1D卷积，类似于Speech2Gesture[10]。我们定义了对抗损失LGAN：

 (12)

其中G被提出为PoseVAE。PoseVAE的总损失可以总结如下。

 (13)

其中MSE, 根, 分别设置为1、1和0.7。

**面部渲染** 我们添加了一个MappingNet来将明确的3DMM系数映射到face-vid2vid[42]的空间中。在训练中，除了在face-vid2vid[42]中使用的损失函数外，我们还在无监督关键点的领域上添加了L1正则化：

 (14)

其中和分别是我们的MappingNet和原始face-vid2vid的运动生成器生成的第n个关键点。L1的权重设置为20，其他损失函数的权重保持与face-vid2vid中相同，它们是在最终生成的图像上计算的。更多细节请参考face-vid2vid[42]。

**B.4.关于对齐系数的更多细节。**

在主论文中，我们展示了在图9中对齐系数的效果。这里，我们提供有关对齐系数的更多细节。通常，对齐系数是变换参数（平移和缩放），用于将任意视频转换和裁剪为对齐的面部视频，以进行深度3D面部重建[5]。PIRenderer[33]的隐式调制包含73维运动系数，包括表情（64）、头部姿势（6）和对齐系数（3）。由于他们的方法专注于视频驱动动画，在测试中对齐系数是已知的。然而，从音频中学习这些系数是困难的，因为对齐和音频之间没有关系。我们也尝试学习并使用我们方法中第一帧的对齐系数（如图9和补充视频所示），然而，产生的头部运动是对齐的且不自然的。我们放弃它以获得更自然的视频结果。因此，我们的运动系数（70）只包含表情（64）和头部姿势（6）。

**C. 补充视频**

我们提供了一个补充视频，包含我们方法和其他相关方法的所有视频结果进行比较，每个组件的消融研究，以及我们方法在不同语言中的更多结果。