arXiv:2403.17694v1 [cs.简历]2024年3月26日

**动画：音频驱动的合成**

华为Wei\*、杨泽军\*、王智胜

腾讯

{huaweiwei, zejunyang, plorywang}@tencent.com

**摘要**在这项研究中，我们提出了一个动画肖像，一个新的框架来生成高质量的动画驱动的音频和参考肖像图像。我们的方法分为两个阶段。首先，我们从音频中提取三维中间表示，并将它们投射到一个二维面部地标序列中。随后，我们采用一个鲁棒扩散模型，加上一个运动模块，将地标序列转换为逼真和时间一致的肖像动画。实验结果表明，羊情画在面部自然性、姿势多样性和视觉质量方面具有优越性，从而提供了增强的感知体验。此外，我们的方法在灵活性和可控性方面显示出相当大的潜力，可以有效地应用于面部运动编辑或面部再现等领域。我们发布代码和模型的权重[https://github.com/Zejun-Yang/AniPortrait.](https://github.com/scutzzj/AniPortrait)

**1介绍**

从音频和静态图像中创建逼真和富有表现力的肖像动画有一系列的应用，从虚拟现实和游戏到数字媒体。然而，制作高质量的动画在视觉上的迷人和保持时间一致性是一个重大的挑战。这种复杂性源于对嘴唇动作、面部表情和头部位置的复杂协调，以创造出视觉上引人注目的效果。

现有的方法往往无法克服这一挑战，主要是因为它们依赖于容量有限的生成器来创建视觉内容，如GANs[[3,](#_bookmark6)[17](#_bookmark1)]、NeRF[[14,](#_bookmark5)[13](#_bookmark2)或基于运动的解码器[[16](#_bookmark3),[8](#_bookmark4)].这些网络表现出有限的泛化能力，并且在生成高质量内容方面往往缺乏稳定性。近年来，扩散模型的出现使得[[2](#_bookmark7),[5](#_bookmark8)[,9]](#_bookmark9)促进了高质量图像的生成。一些研究已经建立在此基础上，通过整合时间模块，使扩散模型能够在创建引人注目的视频方面表现出色。

在扩散模型的进步的基础上，我们引入了安妮肖像画，这是一种新颖的框架，旨在生成由音频和参考图像驱动的高质量的动画肖像画。羊情分为两个不同的阶段。在第一阶段，我们采用基于变压器的模型来提取

1\*平等的贡献。



音频

参考图像

参考姿势图像

姿势引导器

噪声潜伏期

目标姿态图像

Wei等。

从音频输入中得到的三维面部网格和头部姿态序列，随后被投影到一系列二维面部地标中。这个阶段能够捕捉到从音频中微妙的表情和嘴唇的运动，除了与音频的节奏同步的头部运动。在随后的阶段中，我们使用了一个鲁棒扩散模型[[9],](#_bookmark9)与运动模块集成[[4],](#_bookmark10)将面部地标序列转换为一个时间上一致的和逼真的动画肖像。具体来说，我们利用了动画公司的网络架构[[6],](#_bookmark11)该模型利用一个有效的扩散模型，稳定扩散1.5，基于身体运动序列和参考图像生成流体和逼真的视频。特别值得注意的是，我们重新设计了这个网络中的姿态引导器模块。这种修改不仅保持了一个轻量级的设计，而且还提高了产生嘴唇运动的精度。

我们的实验结果表明，羊情画在创造令人印象深刻的面部自然性、不同的姿势和优秀的视觉质量方面的优越性。通过使用三维人脸表示作为中间特征，我们获得了根据需要修改这些表示的灵活性。这种适应性极大地增强了我们的框架在面部运动编辑和面部再现等领域的适用性。

|  |
| --- |
| 头部姿势  …目标姿态图像  项目  目标网格  中性网  音频2姿势  音频2网格      网格  偏移量  **音频2Lmk** |



|  |
| --- |
| 参考Unet  **Lmk2视频**  醋酸乙烯  编码器  …  目标视频  异位去噪    ResNet交叉注意运动模块 |



VAE解码器

去噪Unet



**图1.**对所提出的方法的概述。我们的框架可分为两个阶段。首先，我们从音频中提取三维面部网格和头部姿态，然后将这两个元素投影到二维关键点中。在第二阶段，我们利用扩散模型将二维关键点转换为肖像视频。这两个阶段在我们的框架内同时进行训练。

AniPortrait 3

**2方法**

所提出的框架由两个模块组成，即音频2Lmk和Lmk2视频。前者的设计目的是提取一系列地标，从音频输入中捕捉复杂的面部表情和嘴唇运动。后者利用这个地标序列生成具有时间性的高质量的肖像视频。我们在图中概述了该框架[1](#_bookmark12)并在下面提供更多的细节。

**.12 Audio2Lmk**

让一个1:T=(a1, ..., aT)表示一系列语音片段，我们的目标是预测相应的三维面部网格序列M1:T=(m1, ..., mT在哪里每米t∈ RN ×3，和姿态序列P1:T=(p1, ..., pT)，每个**p**t作为一个表示旋转和平移。

我们使用预先训练过的wav2vec[[1]](#_bookmark13)提取音频功能。该模型具有高度的通用性，能够准确地从音频中识别发音和语调，这在生成真实的面部动画中起着关键作用。通过利用所获得的鲁棒语音特征，我们可以有效地采用一个由两个fc层组成的简单体系结构，将这些特征转换为三维面部网格。我们观察到，这种简单的设计不仅确保了准确性，而且提高了推理过程的效率。

在将音频转换为姿态的任务中，我们采用相同的wav2vec网络作为骨干。然而，我们并不与音频到网格模块共享权重。这是由于姿势与音频中出现的节奏和音调更密切相关，这是一个不同的强调与音频到网格的任务。为了解释以前状态的影响，我们使用了一个变压器[[11]](#_bookmark14)解码器来解码姿态序列。在这个过程中，音频特征被集成到解码器使用交叉注意机制。对于上述两个模块，我们都使用简单的L1损失来训练它们。

在获得网格和姿态序列后，我们利用透视投影将其转换为一个二维的面部地标序列。这些地标随后被用作下一阶段的输入信号。

**2.2 Lmk2Video**

给定一个参考肖像图像，记为Iref，以及一系列面部土地标记表示为L1:T=(l1, ..., lT)每个lt∈ RN×2，我们提出Lmk2视频模块创建了一个时间上一致的肖像动画。此动画将运动与地标序列对齐，并保持与参考图像一致的外观。我们代表肖像动画作为一系列的肖像帧，记为I1:T=(**I**1, ..., IT).

Lmk2视频的网络结构的设计灵感来自于动画中的任何人。我们利用SD1.5作为骨干，结合了一个时间运动模块，有效地将多帧噪声输入转换为一个序列

4魏等。

视频帧。同时，利用反映SD1.5结构的参考网从参考图像中提取外观信息，并将其集成到主干中。这种战略性设计确保了面部标识在整个输出视频中保持一致。与动画不同，我们提高了邮政指南设计的复杂性。最初的版本只包含了几个卷积层，之后，地标特征与主干输入层的潜在元素合并。我们发现，这种基本的设计无法捕捉到嘴唇复杂的运动。因此，我们采用了控制网的控制网[[15]](#_bookmark15)多尺度策略，将相应尺度的地标性特征整合到骨干的不同块中。尽管有了这些增强，我们还是成功地将参数计数保持得相对较低。

我们还引入了一个额外的改进：将参考图像的地标作为一个额外的输入。波后引导器的交叉注意模块促进了参考地标和每一帧的目标地标之间的交互。这一过程为网络提供了额外的线索来理解面部地标和外观之间的相关性，从而帮助生成更精确的运动的肖像动画。

**3实验**

**.13实施细节**

在Audio2Lmk阶段，我们采用wav2vec2.0作为骨干。我们利用媒体管[[7]](#_bookmark16)提取3D网格和6D姿态以进行注释。Audio2Mesh的训练数据来自我们的内部数据集，其中包括来自一个演讲者的近一个小时的高质量语音数据。为了保证媒体管提取的3D网格的稳定性，我们指导演员在保持稳定的头部位置，在整个记录过程中面对摄像机。我们使用HDTF训练音频2的姿势[[18].](#_bookmark17)所有的训练操作都是在一个A100上执行的，使用了Adam优化器，学习速率为1e-5。

在Lmk2视频过程中，我们实现了一个两步训练方法。在最初的步骤中，我们专注于训练主干的二维组件，参考网，和后标指南，而忽略了运动模块。在接下来的步骤中，我们冻结了所有其他组件，并专注于运动模块的训练。我们利用了两个大规模的、高质量的面部视频数据集，VFHQ[[12]](#_bookmark19)和西乐葆-总部[[19]](#_bookmark18)训练模型。所有数据都通过media管道进行处理，以提取二维面部地标。为了提高网络对嘴唇运动的敏感性，我们在渲染二维地标上的姿态图像时，用不同的颜色来区分上唇和下唇。所有图像的大小都被调整为512x512的分辨率。我们使用4个A100gpu进行模型训练，每个步骤都有两天的时间。采用AdamW优化器，学习率一致为1e-5。

**.23结果**

如图所示[2](#_bookmark20)，我们的方法生成了一系列在质量和真实性上都很惊人的动画。我们利用一个中间的三维表示-



参考图像

|  |
| --- |
| 生成的帧 |

音频驱动

参考图像

|  |
| --- |
| 地面真实帧    生成的帧 |

自驱动的

参考图像

|  |
| --- |
| 源帧    生成的帧 |

脸

雷纳克

-ment

AniPortrait 5

文件，可以编辑以操作最终输出。例如，我们可以从一个源中提取地标并改变其ID，从而使我们能够创建一个人脸再现效果。

|  |
| --- |
| 生成的帧  参考图像 |

|  |
| --- |
| 地面真实帧    生成的帧  参考图像 |

|  |
| --- |
| 源帧    生成的帧  参考图像 |

**图2.**我们的方法的结果。

**4.结论和未来的工作**

本研究提出了一种基于扩散模型的肖像动画框架。通过简单地输入音频剪辑和参考图像，该框架能够生成具有平滑嘴唇运动和自然头部运动的肖像视频。利用扩散模型的鲁棒泛化能力，由该框架创建的动画显示了令人印象深刻的真实图像质量和令人信服的逼真运动。然而，这种方法需要使用中间的三维表示，而获得大规模、高质量的三维数据的成本相当高。因此，生成的肖像视频中的面部表情和头部姿势都无法逃脱恐怖谷效应。在未来，我们计划遵循EMO的方法[[10]](#_bookmark21)，预测肖像视频直接从音频，以获得更多惊人的生成结果。

**参考文献**

1.Baevski, A., 周，Y.，穆罕默德， A., Auli，M.： wav2vec 2.0：一个用于语音表征的自我监督学习的框架。神经信息处理系统的研究进展33,12449-12460（2020年）

Wei等。

2.扩散模型在图像合成方面超过了甘斯。神经信息处理系统的研究进展34,8780-8794（2021年）

3.关、J、张、z、周、H、胡、T、王、K、何、d、冯、H、刘、J、丁、E。风格同步：基于风格的生成器中的高保真广义和个性化唇同步。见：IEEE/CVF计算机视觉和模式识别会议论文集。pp.1505– 1515 (2023)

4.郭，Y.，杨， C., 饶的， A., 王、杨，乔、杨、林， D., 奶妈 B.: 动画差异：动画化您的个性化的文本到图像的扩散模型，而不需要特定的调整。arXiv预印本arXiv：2307.04725（2023年）

5.张志明：去噪扩散概率模型。神经信息处理系统的研究进展33,6840-6851（2020年）

6.胡，L.，高，X.，张，P.，孙，K.，张，B.，Bo，L.：为任何人制作动画：人物动画的一致和可控的图像到视频合成。arXiv预印本arXiv：2311.17117（2023年）

7.Lugaresi, C., 唐，纳什，麦克拉纳汉， C., 乌博维贾，E.，海斯，M.，张，F。，张，C。L.，Yong，M。G., 媒体应用：构建感知管道的框架。arXiv预印本arXiv：1906.08172（2019年）

8.马，Y.，张，S.，王，J.，王，X.，张，Y.，邓，Z.：梦谈话：当表达性说话头生成遇到扩散概率模型。arXiv预印本arXiv：2312.09767（2023年）

9.罗姆巴赫，布拉特曼， A., 奥地利动物学家 D., 埃瑟，奥默， B.: 利用潜在扩散模型进行的高分辨率图像合成。见：IEEE/CVF计算机视觉和模式识别会议论文集。pp.10684– 10695 (2022)

10.田，L.，王，Q.，张，B.，波，L.： Emo：情感肖像在弱条件下活着生成表达肖像视频与音频2视频扩散模型。arXiv预印本arXiv：2402.17485（2024年）

11.Vaswani, A., 谢泽尔，帕马尔，谢泽尔，尤斯科雷特，琼斯，戈麦斯， A.N.，Kaiser，L-。我：你所需要的就是注意力。神经信息处理系统的研究进展30（2017）

12.作者：Vfhq：Vfhq：高质量的视频超分辨率数据集和基准。见：IEEE/CVF计算机视觉和模式识别会议论文集。pp.657–666 (2022)

13.叶、Z，何，江，Z，黄，R，黄，J，刘，J，任，Y，尹，X，马、Z，赵，Z：基因脸++：通用和稳定的实时音频驱动3d说话脸生成。arXiv预印本arXiv：2305.00787（2023年）

14.叶，Z，江，Z，任，Y，刘，J，何，赵，Z：基因脸：广义和高保真音频驱动的3d说话脸合成。arXiv预印arXiv：2301.13430

(2023)

15.张志玲：在文本到图像的扩散模型中添加条件控制。见：IEEE/CVF计算机视觉国际会议论文集。pp.3836–3847 (2023)

16.张，W.，村，X，王，X.，张，Y，沈，X，郭，Y，山，Y，王，F.：马鞍人：学习真实的音频驱动的三维运动系数。见：IEEE/CVF计算机视觉和模式识别会议论文集。pp.8652–8661 (2023)

17.张，Z，胡，Z，邓，W，范，C，吕，T，丁，Y：Dinet：变形绘画网络，高分辨率视频的现实面部视觉配音。见：AAAI人工智能会议论文集。卷。37, pp.3543– 3551 (2023)

AniPortrait 7

18.张，Z.，李，李，丁，Y.，范，C.：具有高分辨率视听数据集的流引导的一次性谈话人脸生成。见：IEEE/CVF计算机视觉和模式识别会议论文集。pp.3661–3670 (2021)

19.朱、H.、吴、W.、朱、W.、江、L.、唐、S.、张、L.、刘、Z.、洛伊、C.。C.: 一个大规模的视频面部属性数据集。在：欧洲计算机视觉会议。pp.650–667.施普林格（2022年）