扩散的头部。扩散模型在对话式人脸生成上击败了GANs

米哈乌-斯蒂普乌斯基[1]
michal.stypulkowski@cs.uni.wroc.pl

Konstantinos Vougioukas[2]
k.vougioukas@imperial.ac.uk

何森
senhe752@gmail.com

马切伊-齐安巴[3,4]
maciej.zieba@pwr.edu.pl

Stavros Petridis[2]
sp104@imperial.ac.uk

马贾-潘蒂奇[2]
m.pantic@imperial.ac.uk

[1]弗罗茨瓦夫大学
[2]伦敦帝国学院
[3]弗罗茨瓦夫科学与技术大学
[4]淘宝网

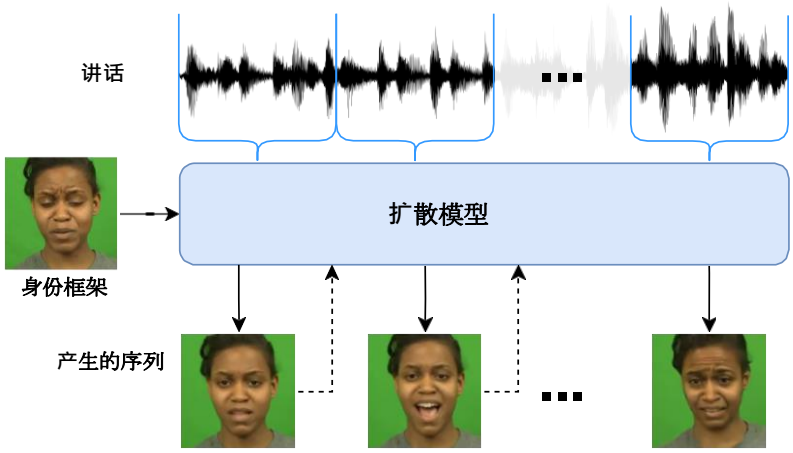

图1.拟议方法的概述。给定一个单一的身份帧和一个包含语音的音频片段，该模型使用扩散模型以自回归的方式对连续的帧进行采样，保留身份，并对嘴唇和头部运动进行建模以匹配音频输入。与其他方法相反，不需要额外的指导。

摘要

在没有其他参考视频的指导下，谈话式人脸生成一直在努力实现头部运动和自然面部表情。最近基于扩散的生成模型的发展允许更真实和稳定的数据合成，它们在图像和视频生成方面的表现已经超过了其他生成模型。在这项工作中，我们提出了一个自动回归的扩散模型，只需要一个身份图像和音频序列就可以生成一个真实的会说话的人头视频。我们的解决方案能够对头部运动、面部表情（如眨眼）产生幻觉，并保留一个给定的背景。我们在两个不同的数据集上评估了我们的模型，在以下方面取得了最先进的结果

他们两个人。[1]

1. 简介

根据语音制作人脸动画有广泛的应用范围，从在连接不良的情况下替代虚拟通话中的视频压缩，到娱乐行业应用的艺术动画，*如*电影、视频游戏和VR体验。然而，鉴于这项任务的众多挑战，目前还没有一个完美的解决方案。到目前为止，现有的方法很难创造出自然的面孔，保持真实的表情和动作，同时在生成过程中仍然需要额外的监督。

深度生成模型在图像和视频生成中不断得到普及并取得令人印象深刻的结果。

[1] 项目页面：https://mstypulkowski.github.io/diffusedheads/。

脸部动画系统的生成任务，已经成为大多数脸部动画系统的事实标准。特别是语音驱动的面部动画系统，它是制作角色动画的一种简单而有效的方法，由于引入了最近的生成模型，如生成对抗网络（GANs）[7]而发生了革命。GANs以能够产生高质量的帧而闻名，同时对生成过程给予很大程度的控制[17, 21, 26]。

尽管GANs具有强大的功能，但其应用于语音驱动的视频合成有几个缺点。首先，GAN训练是出了名的困难，为了达到收敛的目的，需要大量的架构搜索和参数调整。基于GAN的面部动画方法的训练稳定性可以通过使用额外的指导，如面具或驱动帧来指导生成过程来改善。然而，这限制了它们在面部重现方面的应用，降低了它们产生原始头部移动和面部表情的能力。此外，GAN训练往往会导致模式崩溃，即当生成器不能产生覆盖整个数据分布支持的样本，而只能学习生成少数独特的样本时的情况[1]。最后，现有的基于单次GAN的解决方案在生成的视频中存在脸部失真问题，特别是在生成有大量头部运动的视频时。这通常是通过切换到几张照片的方法（*即*使用几帧或一个短片段）或依靠预先训练好的人脸验证模型作为保持身份一致性的神谕来解决的。

我们解决了上述所有问题，提出了Dif- fused Heads--
一种基于框架的扩散模型，可以产生真实的视频，只需要一个身份帧和一个语音记录。生成的头像以一种自然的表现方式移动和行为，同时仍然保留了主体的身份和可信的唇语同步。与最近的大多数方法[3,15,24, 29,35,40,44-
46]相比，我们使用去噪融合概率模型[12,19]，该模型利用变异方法而不是对抗性训练，不需要稳定多个判别器。为了消除看起来不自然的序列的概率，我们引入了运动帧（见第4.2节），经常性地指导视频创作。为了保持语音和生成的帧之间的一致性，我们假设使用通过我们新颖的连接方法注入模型的运动au-dio embeddings。最后，我们没有使用预先训练好的神谕模型，而是对损失函数进行了简单的修改，以保持嘴唇运动的一致性。

我们工作的贡献总结如下。

1. 据我们所知，我们提出了第一个基于扩散模型的对话面生成的解决方案。
2. 我们用运动帧和音频嵌入来充实扩散模型，以保持生成图像的一致性。
3. 我们的方法在泛化方面是稳健的，在身份框架和音频记录的来源上有变化。

2. 相关工作

语音驱动的视频合成问题在文献中得到了很好的定义和探索。它最初是在[43]中研究的，作者在那里发现了声音和视频特征之间有很强的关联性。最早的方法之一是利用隐马尔可夫模型（HMMs）来捕捉视频和语音序列的动态[32, 41, 42]。32]的作者将语音和视频的组合特征表示共同作为全连接马尔可夫模型的状态。在[42]中，作者使用HMMs来估计嘴唇参数的序列。41]的作者提出了一种耦合隐马尔可夫模型（CHMM）的方法，用于视频逼真的语音动画，该方法实现了由与说话人无关的连续语音驱动的逼真面部动画。

遵循机器学习的现代趋势，深度学习方法在基于音频的视频合成领域获得了最有希望的结果。在[37]中，作者提出使用深度学习模型，学习从音素标签输入序列到嘴部运动的任意非线性映射，以准确捕捉自然运动和视觉共鸣效果。在[16]中，对话网络被用来将音频特征转化为特定人物的三维网格。一些方法探索了递归模型的变化[6, 22, 36, 46]。

语音驱动的视频合成的最新方法是基于生成模型的。生成对抗网络（GANs）[7]的变种主要应用于视频生成问题[23, 26, 38]。基于GAN的语音驱动生成方法被引入到[40]中。作者提出了一个端到端的系统，该系统只使用一个人的静止图像和一个包含语音的音频片段，而不依赖手工制作的中间特征，就能生成一个说话的人的视频。他们通过利用时间GAN来实现这一目标，该GAN使用了三个dis-
criminators，专注于实现详细的框架、视听同步和逼真的表情。在[24]中，作者提出在训练过程中加入一个额外的预训练的Lip-
Sync专家，以保持生成视频的一致性。论文[3]介绍了一个三维感知生成网络和一个混合嵌入模块，以保证头部运动的节奏感。文献[45]中预设的模型通过设计一个隐含的低维姿势代码来解决有节奏的头部运动问题，将视听表现模块化。在StyleHEAT[44]中

作者展示了如何利用StyleGAN[17]模型，在语音嵌入的指导下，通过直观或属性编辑来创建会说话的脸。

一些现代的方法利用渲染网络来获得更准确的面部三维表现。在[35]中，作者介绍了一种新型的视频渲染网络和一种动态编程方法，以构建一个时间上共同的、照片上真实的视频。论文[8]的作者提出使用一个音频条件的隐含函数来生成一个动态的神经辐射场，从这个辐射场中，使用体积渲染法同步生成一个与音频信号相对应的高保真的谈话头像视频。论文[29]介绍了一个肖像图像神经渲染器（PIRenderer），它用一个三维可变形脸部模型的参数来控制脸部运动。

ment Learner，连同Dense Warping Field，被用来获得高质量的图像。

去噪扩散概率模型[12]正在获得疯狂的生成能力，并在图像合成[5]和其他引导性图像生成任务[18, 27, 28, 30]中击败了GANs。在无条件的视频生成中，有一些尝试使用了这组模型[9, 13, 14]。在[33]和[11]中研究了带有差异模型的文本驱动的视频合成。

根据我们的知识，在大量文献研究的支持下，还没有直接尝试用扩散模型解决语音驱动的视频合成问题。此外，我们的方法是第一个可以幻化头部运动的一次性方法，不需要演员通过额外的视觉引导输入来驱动运动。手势的真实性与使用明确动作序列的方法相当或更胜一筹。

3. 扩散模型

让我们假设给我们的数据x_0，从数据中取样分布$x_0 \sim q(x_0)$。我们可以定义*前进过程*$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1})$，逐渐增加高斯-数据中的西安噪音。

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t \mathbf{I}) \quad (1)$$

其中T定义了扩散步骤的数量，$\{\beta_t\}_{t=1}^{T}$噪声表从低值开始，随着时间的推移而增加。是t.请注意，$\{\beta_t\}_{t=1}^{T}$是事先知道的，所以整个前进过程是固定的。对于足够大的T和β_t（0，1），x_T变得接近于从各向同性的高斯分布中抽取的样本，*即*$x_T \sim \mathcal{N}(0, \mathbf{I})$。

前进过程的一个有趣的属性是人们只需一个步骤就可以进入其中间状态。就是

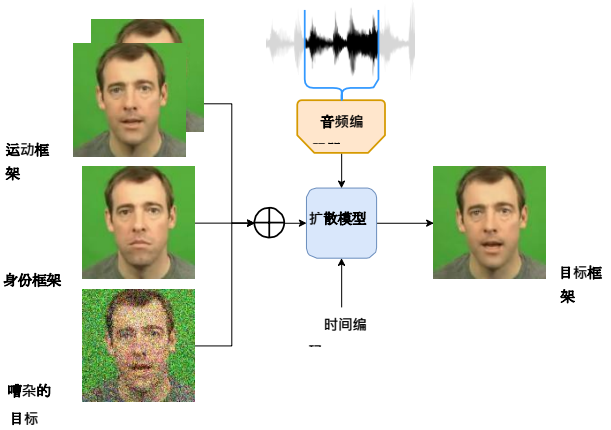

图2.Diffused Heads的训练步骤。我们的模型利用身份帧和运动帧，以及从预训练的音频编码器中提取的音频嵌入，一次学习如何去噪一个帧。身份帧告诉模型感兴趣的脸是什么，而运动帧被用来保留以前的时间步骤的运动。

我们感兴趣的是学习如何对从高斯分布中抽取的样本进行去噪，使其回到数据中。请注意，$q(x_{t-1}|x_t)$取决于整个数据集，因此是难以解决的。直观地说，只有当模型被赋予关于前向扩散过程开始的明确信息时，学习如何将x_T去噪为基础数据点才是可能的。因此，我们可以在x_0上另外建立$q(x_{t-1}|x_t)$的条件，使其具有可操作性。使用贝叶斯定理，我们得到。

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}(x_t, x_0), \tilde{\beta}_t \mathbf{I}) \quad (3)$$

其中。

$$\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \quad (4)$$

$$\tilde{\mu}(x_t, x_0) := \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} x_0 \quad (5)$$

与VAE框架类似，我们定义了近似$q(x_{t-1}|x_t, x_0)$的变量位置。

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (6)$$

如[19]，$\mu_\theta(x_t, t)$和$\Sigma_\theta(x_t, t)$被进一步重新参数化为。

$$\mu(x_t, t) = \frac{1}{\sqrt{\alpha_t}} x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \quad (7)$$

说。

1 -

$$q(x_t|x_0) = N(x_t; \sqrt{\bar{\alpha}}_t x_0, (1 - \bar{\alpha}_t)\mathbf{I}) \quad (2)$$

其中$\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$， 而$\alpha_t = 1 - \beta_t$

。它使我们能够训练

，而v是一个模型，更有效率。

$\bar{\alpha}_t$

$$\Sigma_\theta(x_t, t)) = \exp(v \log \beta_t + (1 - v) \log \beta\text{-t}) \quad (8)$$

其中$E_\theta(x_t, t)$是模型对高斯的预测。

噪声E在得到x_0 的过程中应用于x_t

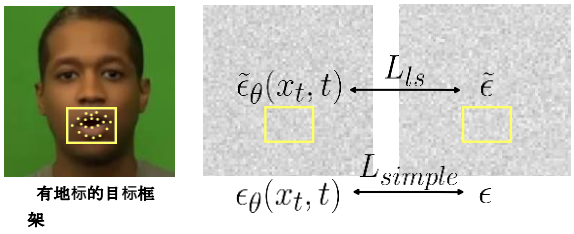

图3.除了在L_{simple}中最小化地面真实噪声E和预测噪声$E_\theta(x_t, t)$之间的L2距离外，我们还利用目标帧的地标来最小化裁剪的地面真实噪声$\tilde{E}$和相应的预测噪声之间的唇语同步损失L_{ls}。面积$\tilde{E}_\theta(x_t, t)$。

模型的额外输出。Nichol & Dhariwal在[19]中提出分别用L_{simple}和L_{vlb}来训练μ_θ和Σ_θ，其中。

$$L_{simple} := \mathbb{E}_{t,x,E}\left[||E - E_\theta(x_t, t)||^2\right] \quad (9)$$

和L_{vlb}是变量下限（VLB），定义为：。

$$L_{vlb} := L_0 + L_1 + \cdots + L_{T-1} + L_T \quad (10)$$

其中

。

$$L_0 := -\log p_\theta(x_0|x_1) \quad (11)$$

$$L_t := D_{KL}(q(x_t|x_{t+1}, x_0) \parallel p_\theta(x_t|x_{t+1})) \quad (12)$$

对于$t \in \{1, \ldots, T-1\}$

$$L_T := D_{KL}(q(x_T|x_0) \parallel p_\theta(x_T)) \quad (13)$$

当数据是图像时，L_0是一个离散的高斯分布，如[12]中提出的。L_T被省略了，因为q没有可训练的参数，$p_\theta(x_T)$是一个高斯先验。所有其他条款都是两个高斯分布之间的Kullback-Leibler发散，可以用封闭形式写出来。

在实践中，一个具有跳过连接的二维UNet[31]，以及注意力层被用来作为一个骨干，以预测两个噪声$E_\theta(x_t, t)$和方差$\Sigma_\theta(x_t, t)$。有关信息时间步数t是使用相应的时间嵌入注入的$\psi(t)$和分组归一化（GN）。

$$h_{s+1} = t_s \, \mathrm{GN}(h_s) + t_b \quad (14)$$

其中h_s和h_{s+1}是UNet的连续隐藏状态。和$(t_s, t_b) =$ MLP$(\psi(t))$，其中MLP是一个由线性层组成的浅层神经网络。

音频编码器。为了达到更平滑和更有表现力的结果，我们通过引入运动帧（第4.2节）和音频运动嵌入（第4.3节）来注入关于过去运动和未来表达的额外信息。此外，一个额外的唇部同步损失（第4.4节）被去掉，以迫使模型更加关注嘴部区域。

4.1. 培训

我们训练一个扩散模型来学习从视频中提取的帧的分布。训练过程如图2所示。我们随机抽取一个视频$X = \{x^{(1)}, \ldots, x^{(n)}\}$，然后从$X$中抽取一帧$x^{(k)}$。$n$是帧的总数。除了标准扩散模型的输入，即时间步长$t$和添加了噪声的帧$x_t^{(k)}$（按照公式（2）），为了保持演员的身份，我们将$x^{(k)}$与一个*身份信息*连接起来。

tity frame x_{Id} channel-wise。

$$x^{(k)}_{在, t} := x_t \oplus^{(k)}_c x_{Id} \quad (15)$$

x_{Id}是从X中随机选择的。在训练过程中随机选择身份帧而不是$x^{(0)}$，使模型熟悉更多种类的帧作为输入。因此，生成的稳健性得到了改善。

为了增加时间上的信息，我们把对应的根据视频中的帧数，将音频序列分成等长的小块。然后，使用在LRW[4]数据集上预训练的audio编码器，将音频块编码为音频嵌入物$Y = \{y^{(1)}, \ldots, y^{(n)}\}$。我们提出的音频处理方法的细节可以在第4.3节中找到。

4.2. 运动框架

尽管音频编码器向模型提供了时间信息，但这还不足以生成流畅的视频。为了克服这个问题并预设运动，对于目标帧$x^{(k)}$，我们引入了*运动帧*$x^{(k)}_{运动} = \mathrm{EB}_c(\{x^{(k-m_x)}, \ldots, x^{(k-1)}\})$。其中$m_x$是运动帧的数量，和$\mathrm{EB}_c(.)$是在通道维度上对其所有的参数进行连接操作。在我们的消融研究中（第5.4节），我们发现m_x的最佳值是2。

如果在$x^{(k)}$之前没有足够的帧，那么最自然的选择是将剩余的运动帧填充到$x^{(0)}$的副本。然而，在采样过程中，我们有

4. 方法

Diffused

Heads每次生成一帧，给定一个身份帧，在整个生成过程中保持固定，并使用预先训练好的语音记录嵌入。

除了身份帧之外，我们不能接触到任何地面真实帧。我们也不一定希望生成的视频以身份帧中给出的准确面部表情开始，*例如*，当音频记录以沉默开始，而身份帧中的人已经张开了嘴。因此，为了使模型对样本的不连续性具有鲁棒性，我们利用$\boldsymbol{x}_{Id}$
作为缺失的运动帧的替代。

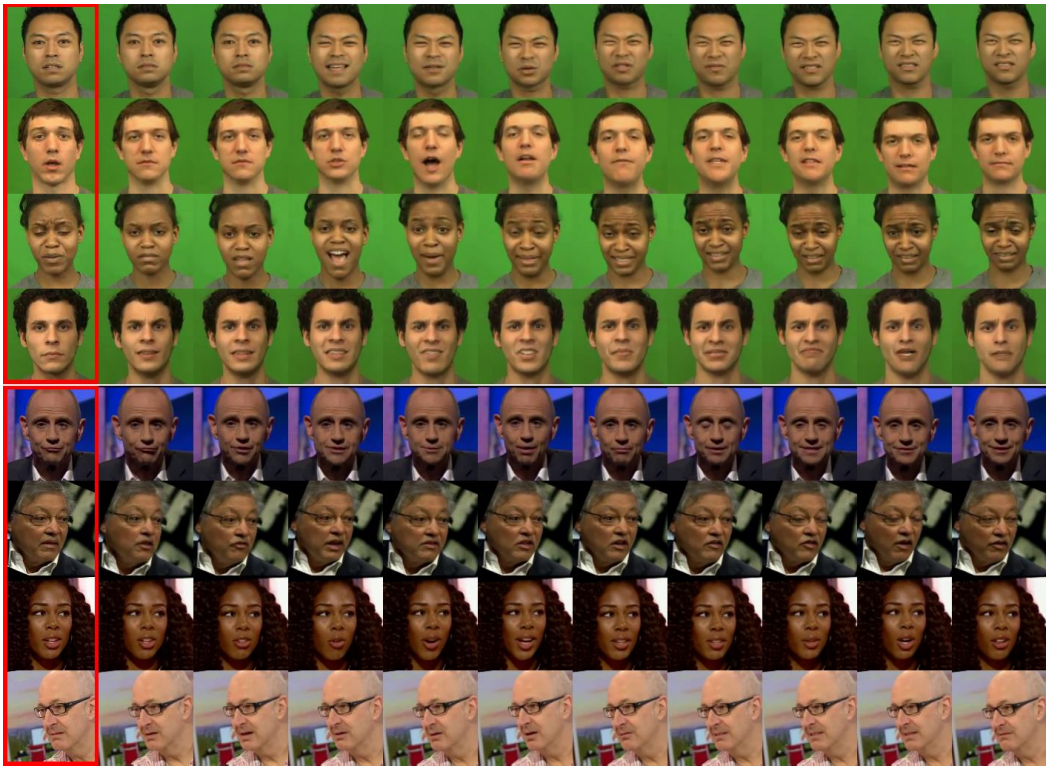

图4.CREMA[2]（顶部）和LRW[4]（底部）数据集的结果。红色边框内的图像是用于生成其余序列的身份帧。

运动帧被添加到方程（15）中，得到模型直接输入的最终形式。

$$x_{在}^{(k)} := x_t^{(k)} \oplus_{cxId} \oplus_c x^{(k)}_{运动} \tag{16}$$

4.3. 语音调节

我们建议从音频嵌入中注入信息 $y^{(k)}$，将方程（14）修改为。

$$h_{s+1} = y_s^{(k)} (t_s \, \mathrm{GN}(h_s) + t_b) + y_b^{(k)} \tag{17}$$

其中$(y_s^{(k)}, y_b^{(k)}) = \mathrm{MLP}(y^{(k)})$。与其他方法[3, 15, 24, 29, 35, 40, 45]不同的是，我们并不在此设置中转移。并用不仅来自时间编码而且来自音频嵌入的信息对UNet的隐藏状态进行调整。我们发现这种方法与其他调节方法相比效果更好，例如在方程（14）的基础上只使用一个额外的比例[25]，以及应用一个多头关注机制，查询是音频嵌入的一个函数[30]。

与在采样过程中只有已经处理过的运动帧相比，我们可以事先获得整个语音记录。为了利用它，我们引入了*运动音频嵌入*，从过去和未来的音频片段中获取信息。 我们把它们定义为一个由以下因素创建的向量串联选定的音频嵌入：$Y^{(k)} = \text{EB}(\{y^{(k-m_y)}, \ldots, y^{(k)}, \ldots, y^{(k+m_y)}\})$，其中$m_y$是来自一方的额外音频嵌入的数量。的我们对$M$的选择的细节 $_y$ 可以在消融中发现在第5.4节中研究。与运动帧类似，如果我们用完了嵌入，我们就用$y^{(k)}$，或者用$y^{(0)}$在开始或$y^{(n)}$在最后。最后，我们用$y^{(k)}$ 在-代替方程（17）中的$y^{(k)}$。

4.4.唇部同步损失

使用任何明确的损失函数来促进生成样本的更好的唇语同步。依靠使用基于预训练的同步或读唇模型的专用感知损失的解决方案，已经有效地提高了唇部识别的准确性[24, 39]，但有两个问题禁止在所提出的模型中使用这种方法。首先，Diffused Heads工作在帧上，而不是序列上，所以基于序列的损失不能被应用。第二，更重要的是，在扩散模型训练期间，目标是预测目标帧上使用的噪声。从预测的噪声返回到初始x_0，这是应用感知损失所需要的，在单一步骤中不够准确，在更多步骤中计算效率低下。

	L W [4]		CREMA [2]	
方法	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑
SDA [40]	**0.76**	23.08	0.70	**23.57**
Wav2Lip [24]	0.73	30.63	-	-
MakeItTalk [46]	0.69	30.38	-	-
PC-AVS [45]	0.71	30.39	-	-
EAMM [15]	0.74	**30.92**	-	-
扩散头（我们的）	0.62	19.11	**0.74**	23.43

表1.与其他方法在选定指标上的比较。

我们引入了一个更简单的解决方案：一个额外的*唇部同步损失*L_{ls}。在训练过程中，我们利用面部地标来裁剪嘴部周围的每一帧，并尽量减少这一区域的噪声预测。

$$L_{ls} = \mathbb{E}_{t,x0,\tilde{E}}\left[||\tilde{E} - \tilde{E}_\theta(x_t, t)||^2\right]。 \quad (18)$$

其中$\tilde{E}$和$\tilde{E}_\theta$，分别表示地面真实和预测噪声的裁剪版本。这个过程在图3中得到了可视化。有了唇部损失，该模型更加关注与音频嵌入的唇部同步，提高了采样视频的整体感知度。我们用一个常数λ_{ls}加权L_{ls}，以利用模型对嘴部区域和帧的其他部分的细节的关注。我们在第5.4节讨论λ_{ls}的选择。

最后的优化目标变成了。

$$L_{simple} + \lambda_{vlb}L_{vlb} + \lambda_{ls}L_{ls} \quad (19)$$

其中L_{simple}和L_{vlb}分别由公式（9）和（10）定义。

4.5.采样

对于采样来说，只需要一个身份帧和从语音记录中提取的音频嵌入点。我们通过初始化$x^{(0)}_{运动}$与身份帧的多个副本来开始视频生成，每一帧都是按照方程（6）中的变异后验定义的扩散模型的标准去噪过程进行采样的。在每一步之后，我们用一个合成的运动帧替换最新的运动帧。$y^{(k)}$遵循与训练期间相同的程序。

生成一个单一的框架需要大量的时间，因为它需要模型对所有的扩散时间步骤$\{1, \ldots, T\}$进行预测。为了加快这一过程，可以使用DDIM[34]或时间步长等方法。在这项工作中，我们使用后者，将采样时间减少了5倍。

在实验中，我们观察到我们的模型在产生突然的头部运动时有时会失败。它逐帧合成序列，任何发生的错误都会在以后的步骤中累积。其中一个相关的问题是，在训练过程中，所有的运动帧都是在训练中产生的。从数据集中提取。同时，在生成过程中，我们使用以前的采样帧，这些采样帧有一些扭曲现象。我们假设，在这种情况下，运动帧和身份帧在提取数据方面是同等重要的。

一个人的属性。

为了迫使模型尽可能多地获取关于

为了尽可能地从身份帧中提取人的外观，我们将每个运动帧转换为灰度。这背后的直觉是，它应该使模型更难提取身份特征（如颜色），同时推动它去寻找运动信息。我们发现这个解决方案在有大量参与者的更复杂的数据集上运行良好。

5. 实验

我们在最常用的对话式人脸生成数据集上评估了Diffused Heads。CREMA[2]和LRW[4]。我们将我们的方法与目前最先进的有引导[15,24,45,46]和无姿势引导[40]的视频合成方法进行定性和定量的比较。为了体验我们成果的全部质量，我们强烈鼓励读者观看补充材料中生成的视频。我们计划发布我们的代码供公众使用。

5.1. 实施细节

我们的模型是在128x128分辨率的视频上训练的。我们使用与[5]中提出的相同的UNet[31]架构，并在第4.3节中解释了音频调节。我们使用256-512-768通道作为输入块。对于每个区块，我们使用2个ResNet[10]层。在实验的早期阶段，我们发现增加更多的注意层会使生成的帧的质量恶化。出于这个原因，我们只在中间区块使用一个具有4个头和64个头通道的注意层。

5.2. 定性结果

我们在图4中介绍了CREMA和LRW的定性结果。视频可以在补充材料中找到。Diffused Heads生成的视频很难与真实的视频区分开来。脸部有自然的表情、眨眼和面无表情。该模型能够保留

方法	分数
PC-AVS [45]	34.95%
扩散头（我们的）	68.72%
真实视频	64.00%

表2.对LRW[4]数据集的图灵测试。每个方法的10个视频和10个真实的视频（共30个）被展示给140人。他们被要求对样本的真实与否进行投票。分数表明视频在参与者看来的真实程度。

帧之间的平滑运动和从一个给定的输入帧的身份。几乎没有任何人工痕迹，而诸如头发或眼镜等困难物体也能准确地生成。通常情况下，Diffused Head在从侧面看人的挑战性视频中表现良好。

需要注意的是，我们的模型没有受到模式崩溃的影响。图4中呈现的结果并不是偷梁换柱的。为了证明我们的说法，我们在项目的网站上分享了整个生成的测试集。

5.3. 量化结果

我们将Diffused Heads与其他方法进行比较。SDA [40], Wav2Lip [24], MakeItTalk [46], PC-AVS [45], 和EAMM [15]。为了进行定量评估，我们从CREMA和LRW数据集的测试分片中生成给定第一帧和音频序列的视频。我们测量SSIM和PSNR以了解结果与参考序列的匹配程度。结果可以在表1中找到。

大多数用于比较的方法都使用了额外的输入来指导生成过程。与[3, 40]类似，我们不提供任何东西，只提供一个单一的框架和音频，允许模型生成任何它想要的东西。由于这个原因，我们合成的视频与参考视频不一致，并且在标准指标中得到更差的衡量。此外，正如[20]中所解释的，PSNR有利于模糊的图像，在我们的任务中不是一个完美的指标，尽管它被普遍使用。

然而，我们认为，我们的结果在视觉数据合成中最重要的指标--人类的感知中获胜。为了证明这一点，我们对140名参与者进行了图灵测试。我们从目前最先进的方法PC-AVS生成的LRW数据集中挑选了10个测试视频，10个来自Diffused Heads，还有10个是在我们的网站上。[2]10个来自Diffused Heads，以及10个真实视频。我们将其发送给来自不同背景和国家的男性和女性。他们中的每一个人在看完30个视频后都被要求投票决定其是否真实。Diffused Heads接近于真实的视频，将其他方法远远甩在后面。详细情况见表2。我们的模型比PC-AVS和令人惊讶的地面真实视频都表现得更好。LRW数据是相当抖动的，因为它们和人头一起移动，有时会产生一种合成的错觉。

[2] 我们发现PC-AVS的样本比EAMM的更真实。

图5.0（上）和2（下）运动帧采样的光流和连续帧的平均幅度。

同时，由于在我们的方法中使用了运动帧，生成的序列更加平滑，使其更加真实。

5.4. 消融研究

在早期的实验中，我们调查了一些运动帧对视频质量的影响。我们注意到，不使用任何运动帧会导致几乎随机的面部表情。为了保持运动，我们试验了最多3个运动帧。

对于1个运动帧，我们取得了更好的结果。虽然仍有抖动，但序列在继续前一帧的运动方面是一致的，并保留了特征。对于2个运动帧，画面非常流畅。没有逐帧失真和突然出现的伪影。我们没能训练出3个运动帧的模型。

在图5中，我们显示了在没有任何眨眼的情况下生成的视频在0和2个运动帧之间的比较。在没有任何运动帧的情况下生成的视频，其光流的平均幅度要比2个运动帧的高得多。在所有的脸部区域，它也是均匀的高。这表明在连续的帧之间有更多的随机运动。对于2个运动帧，我们可以发现嘴巴周围的密度很高，这也是我们所期望的行为。比较单个图像，我们也可以看到使用2个运动帧时更稳定。我们没有将使用1个运动框架模型生成的视频纳入到可视化中，因为仅仅通过观察连续的帧很难区分它和2个运动框架之间的区别。我们在补充材料中包括了差异清晰可见的视频。

对于唇部同步损失权重λ_s，我们观察到大于0.5的值会降低生成视频的质量。在从[0, 0.5]范围内用不同的值运行了几次后，我们决定选择0.2，因为结果仍然非常有吸引力，而且唇部同步得到了改善。

最后，运动音频嵌入的数量和是否在运动帧上使用灰度证明是至关重要的。我们在表3中列出了LRW数据集的消融研究的数字结果。利用灰度可以提高每一种运动音频嵌入数量的生成视频的质量。对于后者，最好的值是2。

运动音频嵌入	运动灰度	CPBD ↑	MSE ↓	SSIM ↑	PSNR ↑	LMD ↓	WER ↓
0		0.0857	1194	0.5767	18.2912	3.0282	0.93
		0.0858	1148	0.5890	18.5003	2.8471	0.93
1		0.0872	1228	0.5782	18.2309	2.8786	0.80
		0.0856	1131	0.5996	18.6589	2.6705	0.81
2		0.0831	1275	0.5658	17.9912	3.0214	**0.72**
		0.0925	**1025**	**0.6225**	**19.1072**	**2.5297**	0.77
3		0.0882	1266	0.5678	17.9945	2.9253	0.74
		0.0882	1184	0.5851	18.4350	2.7590	0.75

表3.对LRW[4]数据集的消融研究。

图6.跨数据集、跨性别、跨语言和跨领域生成的结果。录音是（从头开始）。英语女性，韩国女性，德语男性，以及英语男性。前两行是用CREMA[2]训练的模型生成的，后两行是用LRW[4]的模型。前三行使用了来自AVSpeech的音频，而最后一行，我们使用了自定义的图像和录音。

表明模型有多好，我们在选择最佳模型时主要依靠人类的判断。

我们注意到，在运动帧上使用灰度并不能帮助减少数据集的多样性，例如CREMA。它由91个演员的视频组成，使得对新面孔的归纳更加困难。出于这个原因，使用RGB运动帧可以让模型从身份和运动帧中获取更多的身份信息。

5.5. 归纳

深度学习的主要挑战之一是模型对未见过的数据进行良好的泛化的能力。我们进行了实验来显示Diffused Heads在这种方式下的鲁棒性，结果可以在图6中找到。我们表明，当从不同来源获得部分甚至全部输入时，该模型表现良好。

我们调查了我们在CREMA上训练的模型在CREMA的身份框架上的行为，该身份框架由一个男演员和AVSpeech的两个不同的音频记录组成：一个女性说英语，一个女性说韩语。同样地，我们用LRW模型对一个女性身份进行了测试。

画面和来自AVSpeech的一个男性德国人的音频序列。作为最后的测试，我们使用DALL-E 2[28]合成的图像和我们自己录制的语音，生成了一个会说话的头像视频。

结果显示，Diffused Heads在来自训练分布之外的数据上表现良好。生成的框架看起来很舒服，嘴唇的动作和面部的表情看起来很自然。令人惊讶的是，我们的模型甚至能够成功地处理化身的图像，尽管它在训练期间只得到了人脸。

5.6. 限制条件

尽管Diffused Heads实现了最先进的性能，但它仍然有一些局限性。我们的方法的主要挑战是生成视频的长度。由于我们没有为头部运动提供任何额外的姿势输入或视觉指导，并且模型会自动生成帧，因此它无法保持较长序列的初始质量。此外，与其他生成性模型相比，扩散模型存在生成时间长的问题。目前，不可能使用我们的方法

尽管它在理论上适合于实时应用。用于人脸生成任务的新指标也是一个开放的研究问题。

6. 结论

在这项工作中，我们提出了Diffused Heads：一种基于框架的说话脸部生成方法。为了合成一个人类难以区分的视频，它只需要一个身份帧和一个包含语音的音频序列。我们在两个具有不同复杂程度的数据集上评估了我们的方法，在这两个数据集上取得了最先进的结果。我们对140名参与者进行了图灵测试，显示我们的结果与地面真实视频没有区别，从而支持了这一说法。

参考文献

[1] Martin Arjovsky和Léon Bottou.Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.2

[2] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma.Crema-d：众包的情感多模态演员数据集。*IEEE情感计算交易*，5（4）：377-390，2014。5, 6, 8

[3] 陈乐乐，崔国峰，刘思龙，李忠，寇子义，徐毅，和徐晨亮。用有节奏的头部运动生成说话的头。在*欧洲计算机视觉会议上，sion*，第35-51页。Springer, 2020.2, 5, 7

[4] Joon Son Chung和Andrew Zisserman。野外的读唇术.在*亚洲计算机视觉会议上*，第87-103页。Springer, 2016.4, 5, 6, 7, 8

[5] Prafulla Dhariwal 和 Alexander Nichol.扩散模型在图像合成上击败甘斯。*Advances in Neural Informa- tion Processing Systems*, 34:8780-8794, 2021.3, 6

[6] Bo Fan, Lijuan Wang, Frank K Soong, and Lei Xie.带有深度双向LSTM的照片-真实的说话头。In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing（ICASSP）*, pages 4884-4888.IEEE, 2015.2

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio.Generative adversarial networks.*Communications of the ACM*, 63(11):139-144, 2020.2

[8] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang.Ad-nerf: 音频驱动的说话头合成神经场。在*IEEE/CVF国际计算机视觉会议论文集 sion*，第5784-5794页，2021年。3

[9] William Harvey, Saeid Naderiparizi, Vaden Masrani, Chris- tian Weilbach, and Frank Wood.长视频的灵活扩散建模。*arXiv预印本arXiv:2205.11495*，2022。3

[10] 何开明，张翔宇，任少卿，和孙健.深度残差学习用于图像识别.*Computer-科学》*，2015年。6

[11] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben

Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video:高清晰度视频生成与扩散模 els. *arXiv预印arXiv:2210.02303*, 2022.3

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel.去噪diffusion概率模型。*Advances in Neural Information Processing Systems*, 33:6840-6851, 2020.2, 3, 4

[13] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet.Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.3

[14] Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi.用于视频预测的扩散模型和填充。*arXiv预印本arXiv:2206.07696*，2022。3

[15] 纪新亚，周航，王开元，吴倩仪，吴伟，徐峰，曹勋。Eamm: *arXiv preprint arXiv:2205.15278*, 2022年，通过基于音频的情感感知运动模型的单次情感说话脸。2, 3, 5, 6, 7

[16] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen.通过对姿势和情感的端到端联合学习实现音频驱动的面部动画。*ACM Transactions on Graphics（TOG）*，36（4）：1-12，2017。2

[17] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila.分析和改进stylegan的图像质量。在*IEEE/CVF计算机视觉和模式会议论文集识别*，第8110-8119页，2020。2, 3

[18] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen.Glide:使用文本指导的扩散模型实现逼真的图像生成和编辑。*arXiv预印本 arXiv:2112.10741*, 2021.3

[19] Alexander Quinn Nichol and Prafulla Dhariwal.改进的去噪扩散概率模型.在*国际机器学习会议上*，第8162-8171页。PMLR, 2021.2, 3, 4

[20] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla.Nerfies:可变形的神经辐射场。在*IEEE/CVF计算机视觉国际会议论文集*，第5865-5874页，2021年。7

[21] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski.Styleclip:文本驱动的stylegan图像的操作。在*IEEE/CVF国际计算机视觉会议（ICCV）论文集中*，第2085-2094，2021年10月。2

[22] Hai X Pham, Samuel Cheung, and Vladimir Pavlovic.具有隐性情感意识的语音驱动3D面部动画：一种深度学习方法。In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 80-88, 2017.2

[23] Hai X Pham, Yuting Wang, and Vladimir Pavlovic.Generative adversarial talking head:用一个弱监督的神经网络将肖像带入生活。*arXiv预印本 arXiv:1803.07716*, 2018.2

[24] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar.一个唇部同步专家就可以在野外进行语音到唇部的生成。在*第28届ACM国际多媒体会议论文集中*，第484-492页，2020。2, 5, 6, 7

[25] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn.扩散自动编码器:实现有意义和可解码的表示。在*IEEE/CVF计算机视觉和模式识别会议上*，第10619-10629页，2022年。5

[26] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer.Ganimation:来自单一图像的解剖学感知的面部动画。在*欧洲计算机视觉会议论文集（ECCV）*，第818-833页，2018。2

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision.在*国际机器学习会议上*，第8748-8763页。PMLR, 2021.3

[28] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen.*ArXiv preprint arXiv:2204.06125*, 2022.3, 8

[29] 任玉瑞，李革，陈元琪，李轶群，刘山。Pirenderer:通过语义神经渲染的可控肖像图像生成。在*IEEE/CVF计算机视觉国际会议论文集中*，第13759-13759页。13768, 2021.2, 3, 5

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer.用潜伏扩散模型进行高分辨率图像合成。*在IEEE/CVF计算机视觉和模式会议论文集识别*，10684-10695页，2022。3, 5

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox.U-net:用于生物医学图像分割的卷积网络。在*国际会议上，医学图像通信和计算机辅助干预*，第234-241页。Springer, 2015.4, 6

[32] AD Simons.合成说话的口形的产生- ing头。*Proc. of the Institute of Acoustics*, 1990.2

[33] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Mak-a-video:没有文本-视频数据的文本-视频生成。*arXiv预印本arXiv:2209.14792*，2022。3

[34] Jiaming Song, Chenlin Meng, and Stefano Ermon.去噪扩散隐含模型。*arXiv预印本arXiv:2010.02502*, 2020。6

[35] 宋林森，吴韦恩，陈倩，何然，陈变洛。大家都在说话。让我想说就说。*IEEE Transactions on Information Forensics and Se- curity*, 17:585-598, 2022.2, 3, 5

[36] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman.Synthesizing obama: learn- ing lip sync from audio.*ACM Transactions on Graphics（ToG）*，36（4）：1-13，2017。2

[37] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews.一种用于广义的深度学习方法语音动画。*ACM Transactions on Graphics（TOG）*，36（4）：1-11，2017。2

[38] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz.Mocogan：为视频生成分解运动和内容。在*IEEE计算机视觉和模式识别会议论文集中*，第1526-1535页，2018。2

[39] Konstantinos Vougioukas.*生成现实的人类-haviour*。博士论文，伦敦帝国学院，2022年。5

[40] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic.使用甘斯的真实语音驱动的面部动画。*International Journal of Computer Vision*, 128(5):1398-1413, 2020. 2, 4, 5, 6, 7

[41] 谢磊和刘志强.耦合的hmm方法用于视频逼真的语音动画。*模式识别*, 40(8):2325-2340, 2007.2

[42] Eli Yamamoto, Satoshi Nakamura, and Kiyohiro Shikano.基于隐性马尔可夫模型的嘴唇运动合成。*Speech Communication*, 26(1-2):105-115, 1998.2

[43] Hani Yehia, Philip Rubin, and Eric Vatikiotis-Bateson.声带和面部行为的定量关联。*语音交流*，26(1-2):23-43, 1998.2

[44] 尹飞，张勇，存晓东，曹明登，范艳波，王璇，白青燕，吴宝元，王珏，杨玉九。Styleheat:通过预训练的stylegan生成一次性高分辨率可编辑的谈话脸。*arXiv preprint arXiv:2203.04036*, 2022.2

[45] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu.通过隐式模块化的视听再现生成姿势可控的说话脸。在*IEEE/CVF计算机视觉和模式识别会议上*，第4176-4186页，2021。2, 5, 6, 7

[46] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li.Makelttalk：说话者意识到的谈话头像动画。*ACM Transactions on Graphics（TOG）*，39（6）：1-15，2020。2, 6, 7