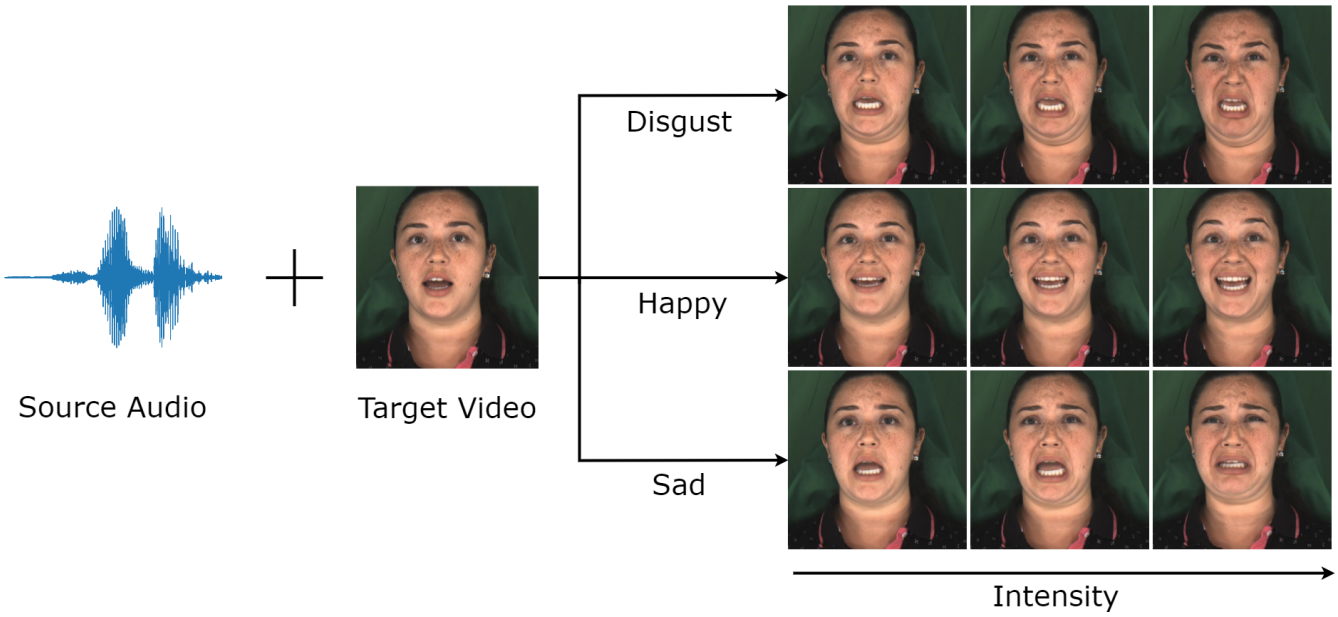
arXiv:2303.00744v1 [cs.CV] 1 Mar 2023

READ Avatars：逼真且可控情绪的音频驱动虚拟形象

Jack Saunders* 和 Vinay Namboodiri

巴斯大学计算机科学系，{jrs68，vpn22}@bath.ac.uk



我们的方法 **READ Avatars** 能够根据音频和参考视频生成具有照片级真实感且唇形同步的视频，并可控制情绪。同一段音频可用于生成具有多种情绪的视频。每种情绪的强度都可以直接指定，从而对输出进行精细控制。

摘要

我们提出了 READ Avatars，这是一种基于 3D 的方法，可生成由音频输入驱动、且能直接精细控制情绪的二维虚拟形象。由于音频到表情的映射具有多对多的特性，以往的方法无法实现逼真的动画。我们通过在音频到表情的生成过程中引入对抗损失来缓解这一问题。这消除了基于回归的模型产生的平滑效应，有助于提升生成的虚拟形象的真实感和表现力。此外，我们指出，音频应直接用于在生成口腔内部时实现了这一点，而其他三维方法并未尝试这样做。我们通过音频条件神经纹理解决这一问题，这种纹理与分辨率无关。为评估我们方法的性能，我们进行了定量和定性实验，包括用户研究。我们还提出了一种新指标，用于比较生成的数字形象对演员情绪的重建效果。结果表明，我们的方法在多项指标上均优于最先进的音频驱动数字形象生成方法。[1]

[1]演示视频可在此处查看：https://youtu.be/QSyMl3vVOpA

1 引言

在影视配音、电子游戏和照片级真实感视频助手等多种应用中，生成逼真的说话头部视频是一项备受期待的能力。尽管这一领域已取得显著进展 [3,8,12,15,16,21,25,28,30–32,36,39,42]„ 但现有方法大多要么使用二维模型实现低质量但准确的唇形同步 [3, 25, 30, 36] ，要么使用三维模型实现高质量但不稳定的唇形同步 [15, 16, 21, 31, 32, 39]。我们推测，有两个关键因素阻碍了兼具高质量和唇形同步能力的模型的发展。第一，音频到表情的映射是一种多对多映射。给定的音频可能对应多种唇形，而相同的唇形也可能产生不同的音频，这是由喉等因素造成的。第二，虽然三维模型通过引入强先验提升了视觉质量，但它们难以表现复杂的唇形，也无法对口腔内部建模（见图 2）。

此外，为了生成更逼真、更可信的视频，并让用户能够在一定程度上控制输出，理想情况下还应引入情绪等额外信号。少数先前的工作尝试过这一点 [12, 13, 15, 24, 37]。这些方法通常要么将情绪视为离散类别 [15, 37], ，这样虽然能够进行语义控制，却缺乏细粒度；要么学习情绪的潜在编码 [12, 13, 24] ，这样可以实现细粒度控制，但不具备语义性，而且需要从其他来源（视频或音频）中选取情绪。

本文提出 READ Avatars，这是一种生成说话头部视频的方法，能够直接、细粒度地控制情绪，同时实现高水平的唇形同步、情绪清晰度和视觉质量。我们采用基于三维的方法，使用可变形模型[20]作为面部的中间表示，并通过延迟神经渲染 [32] 实现高视觉质量。针对上述导致三维模型唇形同步效果不佳的问题，我们提出了两个新颖的组件。首先，我们在音频到表情生成器中加入对抗损失，以缓解多对多映射问题。其次，我们通过以音频为条件的神经纹理，克服使用可变形模型表示复杂唇形和口腔内部的挑战；该神经纹理基于 SIREN 网络 [29]，能够在网格表面编码音频特征，且不受分辨率限制。

总而言之，我们的贡献如下：

- 一种新颖的神经渲染算法，利用神经纹理直接作用于 UV 坐标，并可根据音频进行条件控制，从而改善口腔内部。

- 将 GAN 损失纳入音频到表情网络，以解决音频到表情生成中的多对多问题，从而改善结果。

- 一种用于评估演员的情绪被捕捉和重建得如何的新指标。

2 相关工作

2.1 无条件音频驱动人脸模型

仅根据音频，或根据音频和参考视频生成口型同步视频的任务，称为无条件音频驱动视频生成，已得到广泛研究，并具有许多实际应用，例如配音和数字化身。无条件模型大致分为两类：使用三维先验的模型，以及不使用三维先验的二维模型。

二维模型： 许多从音频合成说话人头部视频的方法，都是直接在图像或视频域中进行操作 [3,25,30]。这些方法通常采用编码器—解码器架构。ATVG [3] 使用音频控制二维关键点，再利用这些关键点生成视频，并通过注意力机制突出需要编辑的部分。Wav2Lip [25] 根据预训练唇形同步检测网络，最小化音频与生成视频之间的距离，从而显著提高唇形同步准确度。尽管唇形同步效果出色，但视觉质量较差。最近，上下文感知变换器 [30] 被应用于这一问题，并结合音频注入式细化网络，显著提升了视觉质量。然而，迄今为止，所有基于二维的模型都受限于视觉质量。相比之下，我们的三维方法能够生成质量高得多的视频。

三维引导人脸模型： 通过显式三维监督，我们创建了由各种信号驱动的超高质量人脸模型 [15,16,18,21,32,34,39]。这些方法通过使用少量参数（例如可直接控制的三维形变模型 [2,5], ）对底层三维场景进行建模，简化了人脸合成。尽管这些模型具有很高的视觉质量，但由于多对多映射问题以及底层几何结构在唇部表情方面的局限，它们往往缺乏表现力。

木偶驱动方法 [15,16], 和运动模型 [12,28]能够通过使用源演员驱动表情，在一定程度上解决多对多问题。这提供了一种信号——源演员的表情——使其与目标演员之间的对应关系更接近一对一。然而，每次使用模型时都要求拍摄源演员并处理生成的视频，通常并不理想。像我们这样不依赖演员的方法具有显著更强的可扩展性。隐式模型 [7,8,23,43]通过多层感知机偏移量增强几何结构，使唇形更具表现力，但无法解决多对多问题。

同期工作 [31]利用记忆网络同时解决了多对多问题和口腔内部的生成问题。然而，这些方法依赖于复用口腔区域的显式像素，导致最终视频出现抖动。

2.2 情绪可控的音频驱动人脸模型

只有少数研究尝试开发能够显式控制生成的说话人头部视频中情绪等风格属性的模型。

MEAD [37] 提出了一种可控制情绪的音频驱动模型。他们训练了一个网络，将音频映射为地标；另一个网络则将输入图像转换为目标情绪，随后使用基于 UNET 的网络，将上半脸、目标情绪与生成的地标结合起来，生成最终图像。这种方法能够控制情绪及其强度，但由于采用逐帧处理，时间连贯性不足，唇形同步效果也不理想。EVP [13]在 MEAD 的基础上进行了改进，增加了一个情绪解耦网络，用于分离音频中的内容与情绪，并采用基于 vid2vid [38] 的人脸合成网络，生成质量更高且时间一致性更好的视频。然而，唇形同步和情绪清晰度并非总能令人满意。MEAD 和 EVP 与我们的工作最为相似。这些基于地标的模型无法达到三维模型那样高的结果质量，而且由于缺少三维形变模型的强先验，也会出现不自然运动。

Kim 等人 [15] 提出了一种基于 3D 的风格可控模型，能够生成高度逼真的视频。该模型使用风格转换网络，将源演员的动画风格转换为目标演员的动画风格。然而，这需要为每种情绪训练一个神经渲染器，并为每对情绪训练一个风格转换网络，因此面对多种风格时很快就会变得难以实现。类似地，EAMM [12] 使用一名源演员来生成情绪风格，并使用运动模型生成唇部运动。该方法无法对口腔区域的情绪进行建模，而且由于情绪风格直接复制自源演员，生成结果在目标演员身上可能显得不自然。NED [24] 也通过三维形变模型参数空间中的情绪操控网络，有效地操控源演员的情绪。该方法能够对情绪进行语义层面的控制，但由于模型的通用性导致情绪风格不匹配，也会产生一些不自然运动。

3 方法

我们的方法包含三个阶段：将三维可变形模型（3DMM）拟合到输入视频上（第 3.1 节），通过对抗训练从音频生成可变形模型参数（第 3.2 节），以及训练音频条件延迟神经渲染器，以生成最终逼真的视频输出（第 3.3 节）。这些步骤如图1所示。

在第一阶段，我们使用经过扩展的 Face2Face 单目重建算法 [34], 将三维可变形模型[2, 19, 20, 44] 拟合到输入视频上，并对算法进行修改，为双眼都加入眨眼混合形状。我们使用 Neural Head Avatars [7] 提供的实现。在第二阶段，我们训练一个受 Pix2Pix [11] 启发的神经网络，通过对抗训练从音频生成可变形模型参数。这样，即使在与音频相关性不强的区域，我们也能生成逼真的动画序列。在此阶段，我们引入细粒度情感标签。在最后阶段，我们训练一个音频条件延迟神经渲染器 [33]。我们的模型使用 SIREN 多层感知机 [29] ，将 UV 坐标直接映射为纹理特征，取代以往工作中的学习式神经纹理 [33], ，从而大大简化音频条件控制任务。

3.1 单目重建

在我们方法的第一阶段，目标是找到一组低维参数，用于建模视频序列 $V = (V_0, \ldots V_n)$。为此，我们使用 FLAME 模型 [20], ，它通过结合蒙皮关节和混合形状来表示显式三维几何。FLAME 模型可以表示为一个函数 $\mathcal{V}$ ，该函数将形状 β、表情 ψ和关节旋转 θ 的一组参数映射为 5023 个三维顶点：

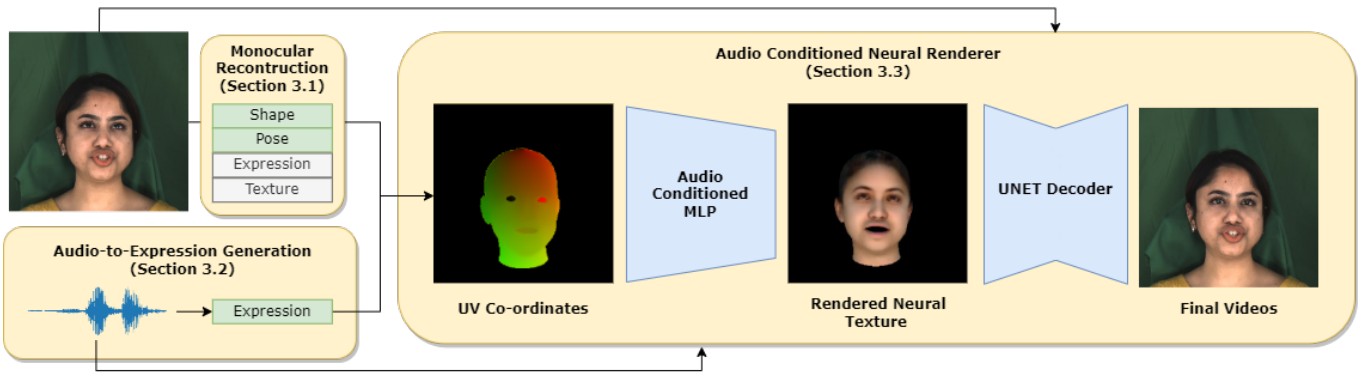


图1：READ Avatars 流程。我们训练了 3 个独立的网络。第一个网络将音频转换为表情（第 3.2 节）。经过光栅化后，第二个基于多层感知机的网络将 UV 坐标转换为神经纹理特征，并以音频为条件（第 3.3 节）。最后，基于 UNET 的网络接收光栅化后的神经纹理和真实视频帧，并输出唇形同步且富有情绪的视频帧。

$$\mathcal{V} : \mathbb{R}^{|\beta|\times|\theta|\times|\psi|} \rightarrow \mathbb{R}^{5023\times 3} \tag{1}$$

我们使用类似的函数，将一组纹理参数 α 映射为基于 UV 的二维纹理。

我们使用全透视相机对渲染过程进行建模，该相机根据以下公式将网格 M 投影到图像平面：

$$\hat{I} = \Pi(M, K, \mathbf{R}, \mathbf{t}) \tag{2}$$

其中 $\mathbf{R}$ 和 $\mathbf{t}$ 分别表示网格在世界空间中的旋转和平移，K 表示相机内参矩阵。我们还将光照建模为远距离漫射光，并使用参数为γ的 3 阶球谐函数。若将 $\pi = (\alpha, \beta, \theta, \psi, \gamma, \mathbf{R}, \mathbf{t}\ K)$ 定义为所有参数的集合，那么我们的目标就是找到最优的$\bar{\pi}$，使其与给定图像最为匹配。

为了将 FLAME 模型拟合到我们的视频上，我们采用 Neural Head Avatars [7], 的跟踪模型，该模型基于 Face2Face [34]。该模型使用 Pytorch3D [26]中的可微渲染来最小化真实帧与渲染帧之间的 L_1距离，并对参数进行统计正则化。

我们假设，对于给定的演员，形状、纹理和光照是固定的，因为我们的数据是在受控条件下采集的。因此，我们可以先估计$\pi_{fix} = (\alpha,\ \gamma,\ K)$，即针对给定主体，在所有视频中保持不变的参数。随后，这些参数将固定用于同一主体所有视频中的所有帧。接着，我们可以逐帧估计 $\pi_{var} = (\theta,\ \psi,\ \mathbf{R},\ \mathbf{t})$。这些参数随后将作为音频到参数生成器的目标。

3.2 音频到参数生成器

我们的方法旨在以音频作为控制信号，驱动照片级真实感虚拟形象。以三维可变形模型为基础的早期方法 [15, 18, 32, 39]使用音频到参数生成器，通过音频或源视频来驱动目标演员。这些方法依赖神经网络及回归损失来生成部分参数。然而，当这些方法将音频映射到表情时，会受到多对多问题的困扰，因为同一段音频可能对应多个同样合理的表情。基于回归的损失意味着，像上半脸运动这样相关性较弱的参数几乎会被完全平均掉，而像嘴唇和下颌这样相关性较强的参数也会被过度平滑。为了解决这些问题，我们提出了一种基于条件生成对抗网络的音频到参数生成器 [6]。我们的模型与 Pix2Pix [11], 类似，结合使用 L_1 损失来处理数据中的低频部分，并使用对抗损失提升真实感。

该网络的输入是一段以梅尔频率倒谱系数表示的音频，$\mathbf{A}$，以及一个显式的情绪标签。MFCC 的窗口大小设定为视频帧率的整数倍。我们使用显式的情绪标注系统，将情绪引入生成参数中。对于 N 种情绪，我们使用一个 $N-1$ 维标签，C，其中中性情绪表示为零向量（无情绪）。每种其他情绪对应一个维度，并按强度缩放，最大强度为 1。这个连续标签能够对情绪进行精细控制。我们将标签沿时间维度分布，



图 2：单目重建方法中典型失败案例的示例。***(左)***：输入帧，***(右)***：重建结果。请注意，FLAME 模型缺乏捕捉某些嘴部形状的表现力，在本例中为“O”形。

以获得 $\mathbf{C} = (C_0, \ldots, C_n)$。该标签与 MFCC 音频特征拼接后，作为音频到表情生成器 $\mathcal{G}_a$ 的输入，生成每一帧的目标参数（$\pi_0, \ldots \pi_n$）。

$$(\pi_0, \ldots \pi_n) = \mathcal{G}_A(\mathbf{A}, \mathbf{C}) \tag{3}$$

判别器是条件式的，它接收真实参数或生成参数，以及音频和情绪标签，并预测给定参数是真实的还是生成的。

生成器和判别器网络均采用编码器-时序-解码器模型，通过一个全连接层将音频特征投影到高维潜在空间，随后接入长短期记忆网络 [10] 和一个全连接解码器，将该潜在空间映射至参数。我们优化以下目标：

$$\mathcal{L} = \mathcal{L}_1 + \lambda_{\text{GAN}}\mathcal{L}_{\text{GAN}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} \tag{4}$$

其中，$\mathcal{L}_1$ 是真实参数与预测参数之间的 ℓ_1 距离，$\mathcal{L}_{\text{GAN}}$ 是对抗损失，$\mathcal{L}_{\text{vel}}$是输出动画速度之间的 ℓ_1 距离。已知速度损失能够提升语音驱动动画 [4]的时序一致性。每个 λ都是相对权重，我们使用 $\lambda_{\text{GAN}} = 0.02$ 和 $\lambda_{\text{vel}} = 100$。

3.3 音频条件化神经渲染器

接下来，我们研究如何逆转参数化模型拟合并生成照片级真实感视频。给定一组与视频$\mathbf{V}$相对应的参数（π_i），我们的目标是尽可能忠实地重现该视频。我们基于神经纹理的思路 [33], ，联合优化图像到图像延迟神经渲染器和定义在 UV 空间中的神经纹理。然而，我们发现 FLAME 模型的表达能力不足以呈现复杂的唇部运动（见图 2），仅靠神经纹理也不足以弥补这一点。此外，FLAME 网格不包含有关口腔内部的信息，包括舌头和牙齿。

为了解决这个问题，我们提出了音频条件化神经纹理。其目标是将音频信息编码到网格表面上，从而能够学习更复杂的唇形和口腔内部。我们用 SIREN 多层感知机 [29] 纹理网络 $\mathcal{T}$取代静态的学习式神经纹理，该网络将光栅化网格的 UV 坐标直接映射为特征向量。这样便无需进行速度较慢且会限制神经纹理分辨率的纹理查找。使用网络还使我们能够轻松地对音频进行条件化，只需将音频特征与 UV 坐标拼接即可。

我们不希望网络中的音频受到情绪或身份的偏置，否则将难以轻松改变生成视频的情绪。为了去除这类信息，我们使用 Wav2Vec2 网络 [1, 40] 的输出。该网络经过预训练，可按 50 fps 预测音素概率。最常见的 50 个音素占音频数据的 99% 以上，因此我们只保留这些音素的特征。接着，我们将这些概率重采样至 60 fps，即视频帧率的两倍。我们截取以目标视频帧为中心、包含 W 帧的窗口，并使用一个小型神经网络 $\mathcal{A}: \mathbb{R}^{2W \times 50}$ 将该窗口内的音频编码为单个向量$\mathbf{a}_{\text{enc}} \in \mathbb{R}^{N_a}$，其中 N_a 表示编码后音频的维度，是一个超参数。随后使用编码后的音频向量对神经纹理进行条件化。

音频编码器由数个全连接层、时间卷积、一个移除时间维度的重塑层以及最终的全连接解码器组成。现在，编码后的音频向量 $\mathbf{a}_{enc}$与光栅化过程中获得的 UV 坐标进行拼接，作为我们的纹理网络的输入。

纹理网络的输出是一幅多通道图像，呈现为带有神经纹理的网格光栅化结果。这种纹理在表面上编码了音频信息。与 [33], 类似，我们使用 16 通道神经纹理，从而能够表示高阶光照效果。随后，这幅光栅化图像会输入一个基于 UNET [27] 的延迟神经渲染器 $\mathcal{R}$，它利用

表1：**与最先进水平的定量比较**。我们使用 FID 比较视觉质量，使用 LSE-D/C [25] 评估唇形同步，并使用我们的指标 A/V-EMD 评估情感重建。我们将结果与 ATVG [3],MEAD [37] 和音频驱动的情感视频肖像 [13] (EVP)进行比较

方法	LSE-C ↑	LSE-D ↓	FID ↓	A-EMD ↓	V-EMD ↓
ATVG [3]	**5.705**	**8.731**	120.040	0.160	0.239
MEAD [37]	4.080	10.569	38.015	0.974	0.113
EVP [13]	4.061	11.514	43.972	0.119	0.126
Ours	4.431	10.157	**13.600**	**0.0686**	**0.093**



图 3：解码器网络的输入由渲染的神经纹理和一帧带有黑色像素边框的真实帧组成。请注意，纹理无法表示口腔内部，但它由解码器生成。

网格上编码的音频特征生成逼真最终帧。

为了解决最终视频中的抖动问题，我们在以目标帧为中心、长度为 W_R 的窗口内，为各帧加入额外的渲染结果。这些额外的渲染结果基于目标帧两侧帧的参数所对应的光栅化 UV 坐标。这样，UNET 解码器的输入就包含 $16+2W_R$ 个通道，从而能够平滑多个帧窗口内的抖动。我们发现，这种方法能有效减少最终视频中的抖动。

为了让我们的视频呈现逼真且时间连贯的背景，我们将纹理网络 V_t 的输出与原始视频帧 V_t 融合。具体而言，我们使用渲染网格的阿尔法通道作为掩码，将前景与背景分离。将前景掩码 α 向外扩展固定数量的像素，得到 α_{exp}。随后，从真实帧中抠除扩展掩码覆盖的区域，并使用前景掩码 α 将渲染网格填入这些像素。经过这一处理，输入由嵌入真实帧中的渲染网格构成，边缘留有一圈可供解码器进行修复的区域。图 3 最直观地展示了这一过程。

$$V_t^* = \alpha \hat{V}_t + (1-\alpha_{\text{exp}})V_t \tag{5}$$

为了训练音频编码器、纹理网络和去耦合神经渲染器，我们端到端地优化以下目标函数：

$$\mathcal{L}(V^*, V) = \lambda_1 \mathcal{L}_1 + \lambda_{\text{VGG}} \mathcal{L}_{\text{VGG}} + \lambda_{\text{GAN}} \mathcal{L}_{\text{GAN}} \tag{6}$$

其中，$\mathcal{L}_1$ 是真实帧与生成帧之间的 ℓ_1 距离，$\mathcal{L}_{VGG}$ 是基于 VGG 的风格损失 [14], ，而 $\mathcal{L}_{\text{GAN}}$ 是对抗损失。超参数 λ_1、λ_{VGG} 和 λ_{GAN} 用于调整各项损失的重要性权重。我们使用 $\lambda_1 = \lambda_{\text{VGG}} = 1$ 和 $\lambda_{\text{GAN}} = 0.01$。

3.4 实现细节

为方法准备数据时，我们首先将每帧裁剪为边长为 256 像素的正方形。具体而言，我们为每帧估算一个带有填充的边界框，然后找出能够覆盖这些边界框并集的最小正方形。再将该正方形调整为所需的分辨率。我们使用 Pytorch 和 Pytorch3D 实现整个流程。所有模型均在一张 NVIDIA RTX3080 显卡上训练。所有网络均使用 Adam [17]进行优化，学习率为 0。0001。渲染器训练 5 个 epoch，耗时约 15 小时；音频到表情生成器训练 10 个 epoch，耗时约 5 小时。所有对抗训练 [22] 均采用 LSGAN 公式。

4 结果

4.1 数据集

我们使用 MEAD 数据集 [37] 进行实验，该数据集包含 60 位演员（男性和女性），他们以 8 种不同的情绪、每种情绪 3 个强度等级朗读 30 个句子，并从多个角度进行录制。在这项工作中，我们仅使用正面摄像头素材。我们遵循 MEAD 中概述的训练测试划分，并针对其中 4 位主体训练模型

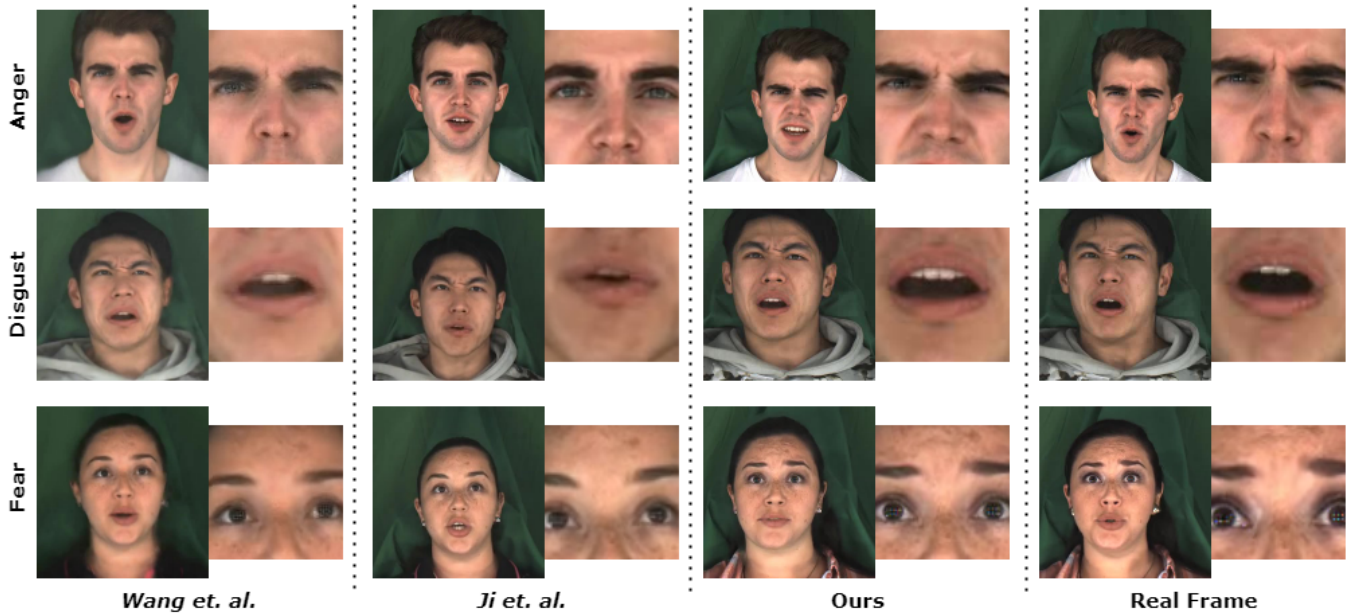


图4：我们将结果与 MEAD [37] 和 EVP [13]进行了比较。我们的结果在视觉质量上远胜 MEAD；放大区域表明，与 EVP 相比，我们的方法生成的情绪更加逼真、准确。

图7展示了我们的方法在多种情绪和强度下取得的部分结果。

4.2 定量指标

为了评估我们方法的性能，我们考察了三个方面：视觉质量、唇形同步和情绪清晰度。

在视觉质量方面，我们使用 Fr´echet Inception 距离（FID）指标来衡量生成帧与真实值的相似度。为了避免因我们的方法使用真实值背景而使结果偏向该方法，我们紧密裁剪面部区域周围的帧。为了衡量嘴唇同步情况，我们采用 wav2lip [25] 中提出的唇音同步误差（LSE）指标。这些指标使用预训练的 syncnet 计算，包括 LSE-D（衡量音频和视频特征之间的最小距离）以及 LSE-C（衡量音频和视频同步的置信度）。要衡量情绪清晰度，仅在帧级别衡量差异并不足够，因为情绪强度会随视频自然变化。因此，我们提出一种衡量情绪清晰度的新指标，用于计算情绪分布之间的差异。该指标基于一个预训练的 EmoNet 模型 [35]。我们预测每帧的效价和唤醒度，并比较生成视频与真实值视频中这些数值的分布。我们使用地球移动距离近似这些分布之间的距离，并分别计算每个主体和每种情绪的距离，取平均值后得到效价情绪均值距离（V-EMD）和唤醒度情绪均值距离（A-EMD）指标。

4.3 与最先进水平的比较

我们将方法与多种最先进的音频驱动虚拟形象模型进行比较。主要比较对象为音频驱动的表情视频生成（EVP）[13] 和多模态情绪感知数据集（MEAD）[37]。MEAD 使用音频到地标的 LSTM、用于将音频驱动的地标转换为任意指定情绪的情绪变换器，以及一个最终的基于 UNET 的模型，根据情绪地标生成输出帧。AudioDrivenEVP 通过设计解耦模型，将音频分离为情绪和内容，从而改进了这一方法；随后，该模型结合地标对齐方法来控制姿态，并利用视频到视频网络，根据地标生成高质量且时间稳定的视频。我们还将方法与

表2：用户研究结果。我们请用户针对四项评判标准，在我们的视频与每种竞争方法之间选择偏好。若用户明显（略微）偏好我们的方法，则结果记为 $++$ （$+$）；若无偏好，则记为 0；若用户明显（略微）偏好另一种方法，则记为 $--$ （$-$）。最右侧列以外的所有数据均以百分比表示。请注意，数据经过四舍五入，因此总和可能不等于 100%。

陈述	$--$	$-$	0	$+$	$++$	mean
我们的方法 > MEAD（口型同步）	1	12	14	28	44	+1.02
我们的方法 > MEAD（视觉质量）	0	3	9	23	65	+1.49
我们的方法 > MEAD（自然度）	1	1	15	23	65	+1.39
我们的方法 > MEAD（情绪）	1	3	21	36	38	+1.06
我们 > EVP（口型同步）	7	16	22	42	11	+0.50
我们 > EVP（视觉质量）	2	22	23	42	11	+0.39
我们 > EVP（自然度）	7	17	18	26	38	+0.49
我们 > EVP（情绪）	4	8	22	26	38	+0.87
我们 > 真实（口型同步）	69	28	3	1	0	−1.64
我们 > 真实（视觉质量）	33	41	23	2	0	−1.05
我们的方法 > 真实（自然度）	53	38	8	1	0	−1.43
我们的方法 > 真实（情绪）	40	38	19	3	1	−1.13

ATVG [3], 进行比较。它是一种基于 2D 的方法，唇形同步效果出色，但视觉质量较差，而且无法编辑情绪。

定量结果： 这些对比的结果如表1所示。我们的结果在视觉质量（FID）和情绪重建（A/V-EMD）方面优于所有竞争方法。虽然我们的方法无法达到无条件 ATVG [3], 的唇部同步质量，但其视觉质量和情绪清晰度要高得多。在唇形同步方面，我们优于两种能够控制情绪的方法：MEAD [37] 和 EVP [13] 。

定性结果： 图4展示了我们与 MEAD [37]和 EVP [13]的结果对比。我们的结果在视觉质量上明显优于 MEAD。与 EVP 相比，我们的方法能够生成更清晰、且更贴近真实视频的情绪，放大区域突显了这一点。尤其可以看出，我们的方法能更好地通过眉毛传达目标情绪。补充视频中还提供了更多结果，进一步展示了我们方法的优势。

用户研究： 为了评估生成虚拟形象的主观质量，我们开展了一项用户研究。我们选取了四位主体，并针对八种情绪中的每一种生成五段视频，共计40段视频。这些视频的背景和姿势参数取自训练集中目标情绪对应的最长视频。我们开展用户研究，将我们的成果与 MEAD [37], EVP [13] 以及真实视频进行比较。我们进行了一项双选强制选择研究，将我们的每段视频与其对应的对照视频配对，后者来自 al-



图5：在神经纹理中加入音频条件控制后，生成帧的质量得到提升，尤其是口腔内部。

替代方案。用户会以随机顺序观看两段视频，并同时看到目标情绪。我们请用户从四个类别中选择哪段视频更好：口型同步、视觉质量、自然度和情绪清晰度。用户可以指出他们是强烈 $++$、较弱地 $+$偏好我们的视频，强烈 $--$ 或较弱地 $-$ 偏好另一段视频，也可以认为两者相同（0）。共有10位用户完成了这项研究。结果如表2所示。我们的方法在所有类别上都显著优于 MEAD。与 EVP 相比，我们的方法在所有类别上也更受青睐。不过，在口型同步和自然质量方面，这种偏好较弱，而在情绪清晰度方面则强得多。用户研究表明，我们的成果仍未达到真实视频的质量，这说明未来工作仍有提升空间。

表3：**消融研究**。我们使用 FID 比较视觉质量，使用 LSE-D/C [25] 比较唇形同步，并使用我们的指标 A/V-EMD 比较情绪重建。我们将完整模型与两个变体进行比较：一个在音频到表情生成器中不使用 GAN 损失，另一个不使用音频条件神经纹理。

方法	LSE-C ↑	LSE-D ↓	FID ↓	A-EMD ↓	V-EMD ↓
Ours	**4.431**	**10.157**	13.600	**0.069**	0.093
我们的方法（无 GAN 损失）	4.047	10.446	**12.587**	0.079	**0.090**
我们的方法（无音频纹理）	4.175	10.398	15.96	0.069	0.96

4.4 消融研究

我们还针对对抗损失和音频条件神经纹理进行了消融研究。该对比的结果如表3所示。加入对抗损失后，口型同步有所改善，但视觉质量略有下降。我们注意到，对抗损失对效价指标影响甚微，但对唤醒度的影响更大。我们推测，这种差异是因为纯回归损失会平滑情绪强度的波动。对于音频条件神经纹理，我们将本方法与静态神经纹理 [33]进行了比较。我们的方法提升了视觉质量和口型同步，并小幅改善了情感重建。正如预期，我们的音频条件控制带来的改善在口腔内部最为明显。这是因为音频使解码器能够区分由相同底层可变形模型几何结构所表示的多个口腔内部。图5展示了这一改进。

5 结论

我们提出了一种能够控制情绪、生成音频驱动虚拟形象的新方法。我们采用基于3D的流程，并在音频到表情生成器中加入对抗损失，同时使用一种由音频条件驱动、分辨率无关的神经纹理。我们的方法缓解了条件音频驱动视频生成中的多对多问题，在唇形同步、视觉质量和情感重建方面均超越了最先进水平，这一点也由我们提出的新指标所凸显。我们的完整解决方案可用于多种应用。

局限性： 使用极端姿态时，我们的模型有时会出现问题（图 6）。此外，由于我们使用参考视频来控制姿态和背景，生成的视频长度受到限制。未来工作将致力于解决任意长度视频的生成问题，

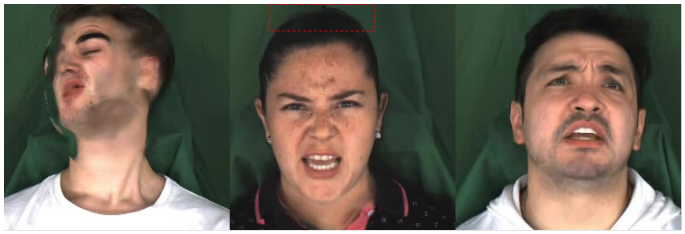


图 6：我们方法的失败案例。**左侧和中间：** 当目标视频的姿态与训练数据显著不同时，会出现伪影。**右侧：** 当跟踪不准确时，我们的模型会生成模糊的结果。

可能会考虑姿态生成。同时，研究能够处理多对多生成的其他模型也很有价值，例如扩散模型 [9, 41]。

伦理影响： 创建合成数字人的能力伴随着被严重滥用的潜在风险。特别是，像我们这样的研究成果可能会被篡改，用来制造令人信服的虚假信息。因此，我们不会向公众开放我们的流程。不过，我们愿意与其他研究人员分享代码。

致谢

本研究部分由 UKRI 资助，资助编号为 EP/S023437/1。

参考文献

[1]Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli。wav2vec 2.0：一种用于语音表征自监督学习的框架。神经信息处理系统进展，33：12449–12460，2020年。[2]Volker Blanz 和 Thomas Vetter。用于合成三维人脸的可变形模型。载于第*26*届计算机图形学与交互技术年会论文集，SIGGRAPH '99，第187–194页，美国，1999年。ACM Press/Addison-Wesley Publishing Co.[3]Lele Chen、Ross K. Maddox、Zhiyao Duan 和 Chenliang Xu。采用动态逐像素损失的分层跨模态说话人脸生成。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2019年6月。[4]Daniel Cudeiro、Timo Bolkart、Cassidy Laidlaw、Anurag Ranjan 和 Michael Black。三维说话风格的捕捉、学习与合成。载于*IEEE*计算机视觉与模式识别会议（*CVPR*）论文集，第10101–10111页，2019年。[5]Bernhard Egger、William A. P. Smith、Ayush Tewari、Stefanie Wuhrer、Michael Zollhoefer、Thabo Beeler、Florian Bernard、Timo Bolkart、Adam Kortylewski、Sami Romdhani、Christian Theobalt、Volker Blanz 和 Thomas Vetter。三维可变形人脸模型——过去、现在与未来。*ACM*图形学汇刊，2020年。[6]Ian J. Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio。生成对抗网络，2014年。[7] Philip-William Grassal、Malte Prinzler、Titus Leistner、Carsten Rother、Matthias Nießner 和 Justus Thies。基于单目RGB视频的神经头部虚拟形象。*arXiv*预印本 *arXiv:2112.01554*，2021年。[8] Yudong Guo、Keyu Chen、Sen Liang、Yongjin Liu、Hujun Bao 和 Juyong Zhang。AD-NeRF：用于说话头部合成的音频驱动神经辐射场。载于*IEEE/CVF*国际计算机视觉会议（*ICCV*），2021年。[9] Jonathan Ho、Ajay Jain 和 Pieter Abbeel。去噪扩散概率模型。*arXiv*预印本 *arxiv:2006.11239*，2020年。

[10]Sepp Hochreiter 和 J¨urgen Schmidhuber。长短期记忆。神经计算，9(8):1735–1780，1997年。[11] Phillip Isola、Jun-Yan Zhu、Tinghui Zhou 和 Alexei A. Efros。使用条件对抗网络进行图像到图像转换，2016年。[12]Xinya Ji、Hang Zhou、Kaisiyuan Wang、Qianyi Wu、Wayne Wu、Feng Xu 和 Xun Cao。Eamm：通过基于音频的情感感知运动模型实现单次情感说话人脸生成。载于*ACM SIGGRAPH 2022*会议论文集，SIGGRAPH '22，美国纽约，2022年。计算机协会。[13] Xinya Ji、Hang Zhou、Kaisiyuan Wang、Wayne Wu、Chen Change Loy、Xun Cao 和 Feng Xu。音频驱动的情感视频肖像。载于*IEEE*计算机视觉与模式识别会议论文集 *(CVPR)*，2021年。[14] Justin Johnson、Alexandre Alahi 和李飞飞。用于实时风格迁移和超分辨率的感知损失。载于欧洲计算机视觉会议，第694–711页。施普林格，2016年。[15] Hyeongwoo Kim、Mohamed Elgharib、Hans-Peter Zoll¨ofer、Michael Seidel、Thabo Beeler、Christian Richardt 和 Christian Theobalt。保留神经风格的视觉配音。*ACM*图形学汇刊 *(TOG)*，38(6):178:1–13，2019年。[16]Hyeongwoo Kim、Pablo Garrido、Ayush Tewari、Weipeng Xu、Justus Thies、Matthias Nießner、Patrick P´erez、Christian Richardt、Michael Zoll¨ofer 和 Christian Theobalt。深度视频肖像。*ACM*图形学汇刊 *(TOG)*，37(4):163，2018 年。[17]Diederik P. Kingma 和 Jimmy Ba。Adam：一种随机优化方法。载于 Yoshua Ben-gio 和 Yann LeCun（编），第*3*届国际学习表征会议（*ICLR 2015*），美国加利福尼亚州圣迭戈，*2015*年*5*月*7*日至*9*日，会议论文集，2015年。[18] Avisek Lahiri、Vivek Kwatra、Christian Frueh、John Lewis 和 Chris Bregler。LipSync3D：利用姿态和光照归一化，从视频中数据高效地学习个性化三维说话人脸。载于*IEEE/CVF*计算机视觉与模式识别会议论文集 *(CVPR)*，2021年。

[19] J. P. Lewis、Ken Anjyo、Taehyun Rhee、Mengjie Zhang、Fred Pighin 和 Zhigang Deng。混合形状面部模型的实践与理论。载于 Syl-vain Lefebvre 和 Michela Spagnuolo（编），*Eurographics 2014 - 最新技术报告*。Eurographics 协会，2014年。

[20] Tianye Li、Timo Bolkart、Michael. J. Black、Hao Li 和 Javier Romero。从四维扫描中学习面部形状和表情模型。*ACM*图形学汇刊（*SIGGRAPH Asia*论文集），36(6):194:1–194:17，2017年。

[21] Yuanxun Lu、Jinxiang Chai 和 Xun Cao。实时语音肖像：实时照片级逼真的说话头部动画。*ACM*图形学汇刊，40(6)，2021年12月。

[22] Xudong Mao、Qing Li、Haoran Xie、Raymond YK Lau、ZhenWang 和Stephen PaulSmolley。最小二乘生成对抗网络。载于*IEEE*国际计算机视觉会议论文集，第2794–2802页，2017年。

[23]Ben Mildenhall、Pratul P. Srinivasan、Matthew Tancik、Jonathan T. Barron、Ravi Ramamoorthi 和 Ren Ng。NeRF：将场景表示为用于视图合成的神经辐射场。载于*ECCV*，2020年。

[24] Foivos Paraperas Papantoniou、Panagiotis P. Fil- ntisis、Petros Maragos 和 Anastasios Roussos。神经情感导演：在“野外”视频中对面部表情进行保留语音的语义控制。发表于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2022年。

[25] K R Prajwal、Rudrabha Mukhopadhyay、Vinay P. Namboodiri 和 C.V. Jawahar。只需一个唇形同步专家，即可实现野外语音驱动唇形生成。发表于第*28*届*ACM*国际多媒体会议论文集，MM ’20，第484–492页，纽约，纽约州，美国，2020年。计算机协会。

[26] Nikhila Ravi、Jeremy Reizenstein、David Novotny、Taylor Gordon、Wan-Yen Lo、Justin Johnson 和 Georgia Gkioxari。使用Pytorch3D加速三维深度学习。*arXiv:2007.08501*，2020年。

[27]Olaf Ronneberger、Philipp Fischer 和 Thomas Brox。U-Net：用于生物医学图像分割的卷积网络。载于 Nassir Navab、Joachim Hornegger、William M. Wells 和 Alejandro F. Frangi 编，医学图像计算与计算机辅助干预——*MICCAI 2015*，第234–241页，沙姆，2015年。施普林格国际出版公司。[28] Aliaksandr Siarohin、St´ephane Lathuili`ere、Sergey Tulyakov、Elisa Ricci 和 Nicu Sebe。用于图像动画的一阶运动模型。载于神经信息处理系统会议（*NeurIPS*），2019年12月。

[29]Vincent Sitzmann、Julien N.P. Martel、Alexan- der W. Bergman、David B. Lindell 和 Gordon Wetzstein。《具有周期性激活函数的隐式神经表示》。载于 *NeurIPS*论文集，2020年。

[30] Yasheng Sun、Hang Zhou、Kaisiyuan Wang、Qianyi Wu、Zhibin Hong、Jingtuo Liu、Errui Ding、Jing- dong Wang、Ziwei Liu 和 Koike Hideki。《借助 Transformer 中的视听上下文挖掘进行掩码唇形同步预测》。载于 *SIGGRAPH Asia 2022*会议论文集，SA ’22，纽约，纽约州，美国，2022年。计算机协会。

[31] Anni Tang、Tianyu He、Xu Tan、Jun Ling、Runnan Li、Sheng Zhao、LiSong 和 Jiang Bian。《记忆是说话人脸生成中的一对多映射缓解器》，2022年。

[32] Justus Thies、Mohamed Elgharib、Ayush Tewari、Christian Theobalt 和 Matthias Nießner。神经语音木偶：音频驱动的人脸重演。*ECCV 2020*，2020年。

[33]Justus Thies、Michael Zollh¨ofer 和 Matthias Nießner。延迟神经渲染：使用神经纹理进行图像合成。*ACM*图形学汇刊（*TOG*），38(4):1–12，2019年。

[34]Justus Thies、Michael Zollh¨ofer、Marc Stamminger、Christian Theobalt 和 Matthias Nießner。Face2face：实时人脸捕捉与 RGB 视频重演。*ACM*通讯，62(1):96–104，2018年12月。

[35] AntoineToisoul、Jean Kossaifi、AdrianBulat、Georgios Tzimiropoulos 和 Maja Pantic。在自然场景条件下，从人脸估计连续效价和唤醒度。自然机器智能，2021年。

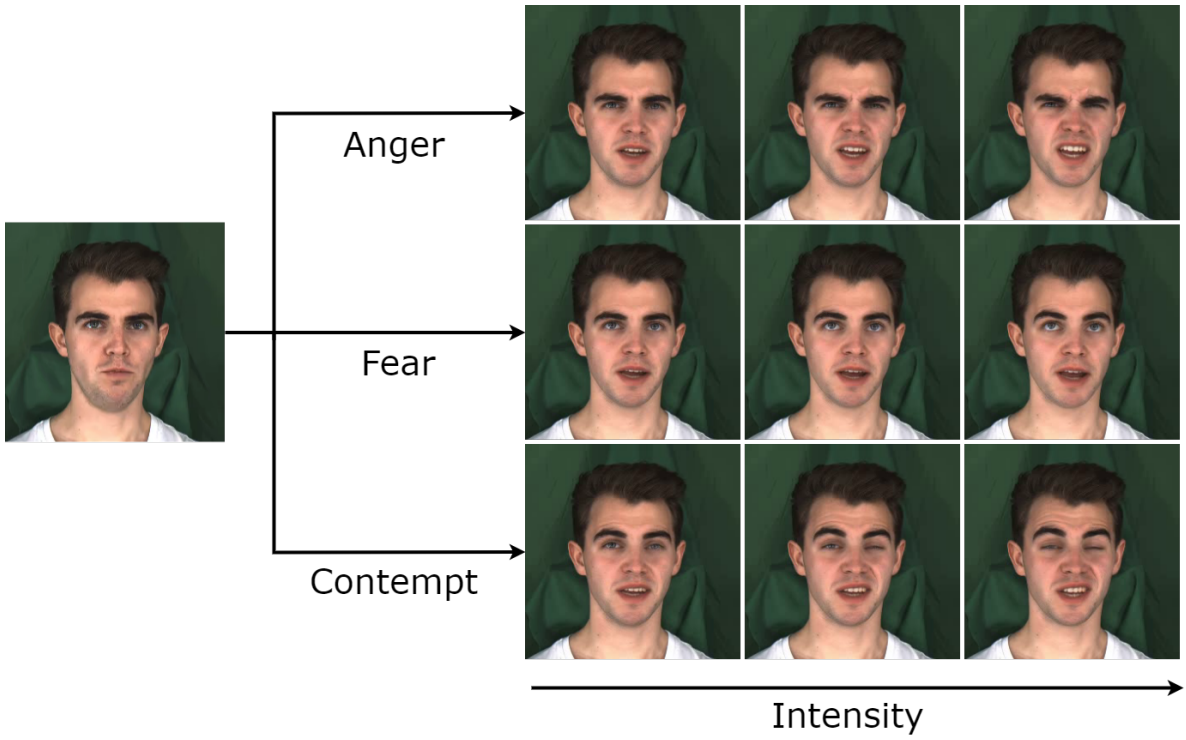
[36]Konstantinos Vougioukas、Stavros Petridis 和 Maja Pantic。使用时序生成对抗网络进行端到端语音驱动人脸动画。*arXiv* 预印本 *arXiv:1805.09313*，2018年。

[37] Kaisiyuan Wang、Qianyi Wu、Linsen Song、Zhuo- qian Yang、Wayne Wu、Chen Qian、Ran He、Yu Qiao 和 Chen Change Loy。MEAD：用于情感说话人脸生成的大规模视听数据集。发表于 $ECCV$，2020年。

[38]Ting-Chun Wang、Ming-Yu Liu、Jun-Yan Zhu、Guilin Liu、Andrew Tao、Jan Kautz 和 Bryan Catanzaro。视频到视频合成。载于神经信息处理系统进展（*NeurIPS*），2018年。

[39] Xin Wen、Miao Wang、Christian Richardt、Ze-Yin Chen 和 Shi-Min Hu。照片级真实感音频驱动的视频肖像。*IEEE* 可视化与计算机图形学汇刊，26(12):3457–3466，2020年。

[40]Qiantong Xu、Alexei Baevski 和 Michael Auli。简单有效的零样本跨语言音素识别。*arXiv* 预印本 *arXiv:2109.11680*，2021年。[41]Ling Yang、Zhilong Zhang、Yang Song、Shenda Hong、Runsheng Xu、Yue Zhao、Yingxia Shao、Wentao Zhang、Bin Cui 和 Ming-Hsuan Yang。扩散模型：方法与应用的综合综述。*arXiv* 预印本 *arXiv:2209.00796*，2022年。[42] Xinwei Yao、Ohad Fried、Kayvon Fatahalian 和 Maneesh Agrawala。利用神经重定向对说话头进行迭代式基于文本的编辑。*ACM*图形学汇刊，40(3)，2021年8月。[43] Yufeng Zheng、Victoria Fern´andez Abrevaya、Mar- cel C. B¨uhler、Xu Chen、Michael J. Black 和 Ot- mar Hilliges。I M Avatar：从视频中生成的隐式可变形头部化身。载于计算机视觉与模式识别（*CVPR*），2022年。[44]M. Zollh¨ofer、J. Thies、P. Garrido、D. Bradley、Thabo Beeler、P. P´erez、M. Stamminger、M. Nießner 和 C. Theobalt。单目三维人脸重建、跟踪及应用的最先进水平。计算机图形论坛，37:523–550，2018年5月。



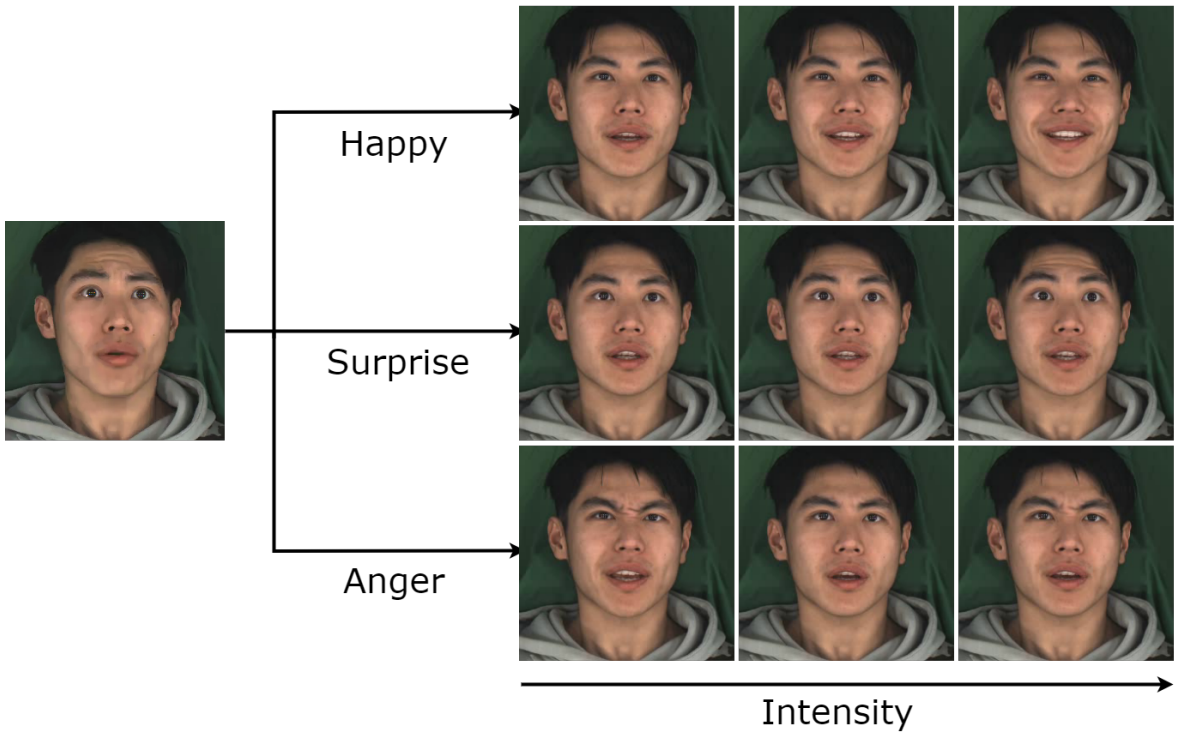


图7：我们的方法支持对多个主体进行精细控制。这里展示了两个主体分别呈现三种情绪和三个强度等级。