HeadGAN: 一次性神经头部合成和编辑

------CVPR2021

Michail Christos Doukas[1,2], Stefanos Zafeiriou[1,2], Viktoriia Sharmanska[1,3]
[1] Imperial College London, UK [2]Huawei Technologies, London, UK [3] University of Sussex, U

图 1：我们提出的 ***HeadGAN*** 方法通过将面部表情和头部姿势从驱动框架完全转移到参考图像来执行重演（b）。 当驱动身份和参考身份重合时（c），它可以用于面部视频压缩和重建。 此外，***HeadGAN*** 可应用于面部表情编辑 (d)、新视图合成 (e) 和面部正面化 (f)。 项目页面：https://michaildoukas.github.io/HeadGAN/

Abstract

最近使用单一参考图像来解决头部再现问题的尝试已经显示出很有希望的结果。然而，它们中的大多数要么在照片真实感方面表现不佳，要么未能满足人物保存问题，要么未能完全转移驱动姿势和表情。我们提出了 ***HeadGAN***，这是一种基于 3D 人脸表征进行合成的新系统，它可以从任何驱动视频中提取，并适用于任何参考图像的面部几何形状，将人物与表情分离开来。我们进一步改善口型运动，利用音频特征作为补充输入。3D 人脸表示使 ***HeadGAN*** 进一步成为一种有效的压缩和重建方法，以及表达和姿态编辑工具。

1. Introduction

视觉数据合成[44, 43]，包括会说话的头部动画[45, 49, 48, 17, 33, 34, 23]是令人兴奋和蓬勃发展的研究领域，在编辑、游戏、社交媒体、虚拟现实、电话会议和虚拟辅助等领域有着无数的应用。在过去的几年里，解决方案主要是由图形学提供的。例如，***Face2Face***[40]方法通过从驱动视频中恢复面部表情并将其覆盖到源帧来实现面部再现。最近一些基于学习的方法[23, 25, 13]寻求解决头部完全再现的问题，其目的是将驱动人的表情和姿势转移到源人物。这种方法的缺点是它们依赖于来自源视频的长视频片段，因为它们训练的是特定于人的模型。

与此同时，人们提出了多种方法来在少量镜头下重现人类头部[45, 43, 48, 17, 34, 14]，其中只有有限数量的参考图像，甚至只有一张。大多数最先进的方法使用面部关键点来指导

合成[49,48,43,34]，通常因为关键点对外观信息进行编码，在再现过程中存在人物保存问题。当源视频的头部几何形状与驱动视频中的人不同时，这个问题就变得更加突出。

在本文中，我们提出了 *HeadGAN*，一种新的一次实现头部动画合成和编辑的方法。我们采用了与大多数现有的少镜头方法不同的方法，使用类似于 PNCC[52]的 3D 人脸表示来条件合成。我们利用了表达和人物解开的先验知识，包含在 ***3D Morphable Models***(3DMMs)中[3,5,4,6]。我们决定使用 **3DMMs** 进行人脸建模，使 *HeadGAN* 具有以下功能:1)运行在每秒 20 帧的实时再现系统;2)有效的面部视频压缩和重建方法;3)面部表情编辑方法；4)新颖的视图合成系统，包括正面化。图 1 展示了我们方法支持的任务。除了 3D 人脸，我们还可以选择将生成过程置于音频信号的语音特征上，使我们的方法能够执行准确的口腔合成，正如我们的自动唇读实验所建议的那样。

我们与最先进的方法[45, 43, 48, 34, 31, 51]进行了广泛的比较，并根据标准 GAN 指标[19, 41]生成了卓越的图像质量和性能，在以下任务中：重建、重建和正面化，甚至与在更大的 *VoxCeleb2*[9]数据集上训练的模型[48]相比也是如此。 最后，我们进行了消融研究，以证明我们系统的每个组件的贡献。

2. Related Work

用于人脸合成的无模型方法。*X2Face*[45]是最早的基于学习的人头动画方法之一，它不依赖于人脸的任何先验知识。在某些情况下,它们的翘曲操作会导致不自然的头部变形，从而导致照片真实感不佳。 *MonkeyNet*[33]是一个更新的深度学习框架，它建议通过驱动视频的关键点检测来推断运动。 然后，使用从参考图像中提取的外观以及运动信息来生成输出。 在后续工作中，一阶运动模型（**FOMM**）[34]显著提高了单幅图像动画的效果。**FOMM** 使用相对关键点位置来保留源的人物,这要求驱动视频第一帧中的对象与源图像中的对象处于相同的姿势。 由于并不总是满足这样的假设，因此不能保证生成样本的头部姿势会跟随驱动员的头部姿势。

基于 *landmark* 的人脸建模和生成。*Bringing Portraits to Life*[1]是为静止图像制作动画的最初尝试之一。在源图像上应用 2D 扭曲以模仿驱动视频中的面部变换。当源头部姿势接近目标图像中出现的姿势并且只需要很小的变形时，它显示出有希望的结果。 *Warp-Guided GANs*[14]是最近的一项工作，它使用 2D 面部标志和 2D 扭曲来为图像设置动画。它需要一张以正面姿势拍摄的表情中性的照片。许多关于头部动画的研究都假设了几个展示设置，其中有少量的参考图像可用。*Zakharov et al.*[49]从参考图像中提取与人物相关的嵌入，并通过自适应实例归一化层(*AdaIN*)[20]将它们注入生成器。在对新人物进行微调后,他们基于图像的方法表现最佳。*Bi-layer Neural Avatars*[48]是另一种利用 *SPADE*[30]的一次性模型，同时在推理过程中以实时速度运行。

早先已经提出使用 *SPADE layers*[30]使生成过程适应源的外观，作为 *few-shot vid2vid model*[43]的一部分，这是一种扩展 *vid2vid*[44]的基于视频的方法。 大多数上述方法无法解决重建中的人物保存问题,因为面部标志允许将来自驱动员的人物相关信息传输到生成的样本中。*MarioNETte*[17]试图通过提出一种 *landmark* 变换方法来解决这个问题，该方法使驱动 landmark 适应参考头部的形状。

由 3D 面孔辅助的头部动画。**3DMM**[3, 5, 4, 6]已被证明对人脸建模非常有效，并已广泛用于驱动人脸合成[40, 23, 35, 39]。拟合 **3DMMs**，可以从目标帧中恢复准确的姿势和表情，以及从参考图像中恢复与人物相关的参数。然后，渲染的 3D 面部用于调节神经网络，完成纹理并填充信息缺失的区域（头发、身体、背景等）。*Deep video portraits* (DVP)[23]和 *Head2Head*[25]是此类由 3D 信息驱动的重建系统的示例。这两种方法都使用源风格的长

视频片段来训练特定于个人的模型。相反，我们提出的方法是通用的，即它可以使用单个参考图像为任何看不见的人执行视频合成。其他方法如 *StyleRig*[38]和 *GIF*[15]使用 3DMM 控制 *StyleGAN*和 *StyleGAN2*[21, 22]但未能保留面部模型未解释的场景部分，如头发、背景。

音频驱动的头部合成。 除了上面讨论的视频驱动技术之外，还有大量文献关注音频驱动的人脸合成[37, 8, 35, 7, 42, 39]。与这些方法有很大不同，我们的系统可以选择使用音频信号来增强嘴巴区域的语音质量和真实感，而姿势和表情则由目标视频片段引导。

3. Methodology

3.1. 三维人脸表示

为了在保留源人物的面部几何形状的同时准确地传递驱动者的表情，我们利用了包含在 3DMM 中的人脸先验知识[3,5,4,6]。给定 T 帧的驱动视频，$y_1 = \{y_t | t = 1, \dots, T\}$，3DMM 拟合阶段产生一系列相机参数$\boldsymbol{c}_{1:T}$和形状参数$p_{1:T}$，其中$p_t = [p_t^{idT}; p_t^{expT}]^T$。也就是说，对于每一帧 t，我们获得两种类型的形状参数：a) 人物相关参数$p_t^{id} \in R^{n_{id}}$，编码面部几何和 b) 表情参数$P_t^{exp} \in R^{n_{exp}}$，表示面部变形。这使得能够从运动引起的形状变形中解开依赖于人物的面部形状属性。我们恢复了非常准确的面部表情，因为我们的 3DMM 拟合阶段依赖于一组密集的 3D 点（大约 1K）。这些点是从带有 *RetinaFace*[12]的帧中回归的，该帧在 WIDER FACE 数据集[47]上进行了预训练。此外，给定源标识y_{ref}的参考图像，我们执行 3DMM 拟合以获得源的形状参数$P_{ref}^{id}, P_{ref}^{exp}$和相机参数$c_{ref}$。有关 3DMM 拟合算法的详细信息，请参阅补充材料。

接下来，对于每一帧 t，我们计算 3D 面部形状（3D 网格）$s_t = [x_1, y_1, z_1, \dots, x_N, y_N, z_N]^T \in R^{3N}$，如$s_t = x + U^{id} P_{ref}^{id} + U^{exp} P_t^{exp}$。 (1) 这里$x \in R^{3N}$是均值形状，$U_{id}$是恒等正交基，$U_{exp}$是 LSFM 可变形模型的表达式正交基[6]。通过构造，这个 3D 形状s_t反映了源的面部结构P_{ref}^{id}和驱动人物的面部表情P_t^{exp}。通过这种方式，我们解决了源人物保存问题。最后，我们使用 3D 形状s_t和相机参数c_t来渲染 3D 人脸表示$x_t = R(s_t, c_t)$。如图 2 所示，这是一个类似于 *PNCC*[52]的 RGB 图像。我们还使用由等式 1 获得的P_{ref}^{exp}和由P_{ref}^{id}得到的相机参数c_{ref}和形状s_{ref}来渲染x_{ref} ，它是从参考图像y_{ref}恢复的 3D 人脸。 在 4.1 中，我们将更详细地讨论 3D 人脸渲染。

总而言之，给定一个驱动视频$\boldsymbol{y}_{1:T}$和一个源图像y_{ref}，数据预处理管道恢复一系列图像$\boldsymbol{x}_{1:T}$，它描绘了从驱动程序中提取的 3D 人脸，并适应了源的面部几何形状，如以及参考图像外部参照的 3D 面。这些面部表征用于通过 *HeadGAN* 的生成器网络来调节图像合成。

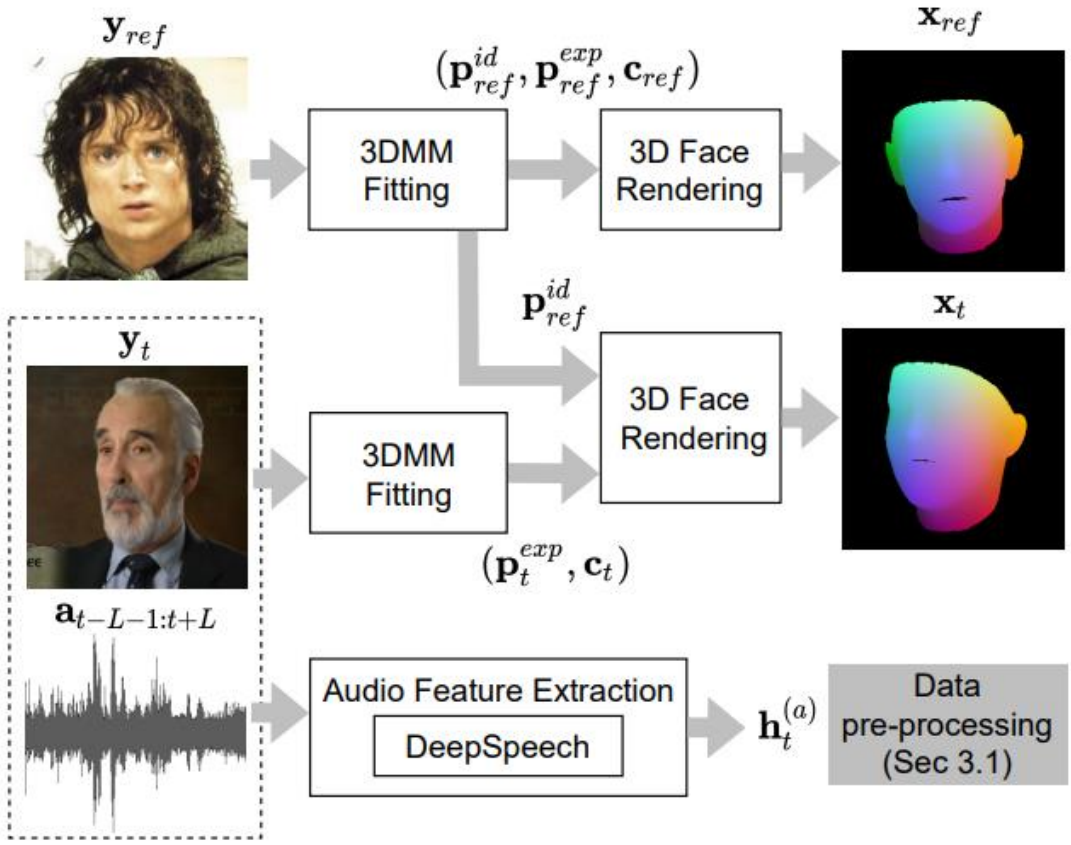

图 2：数据预处理阶段。 在调整身份参数之后，我们恢复并渲染参考图像y_{ref}的 3D 人脸，以及驱动框架y_t。

3.2 音频特征

与之前的一次性头部重建系统相反，我们的方法利用了驱动音频流及其与面部和嘴巴运动的相关性。我们将音频信号分成 T 个部分$\boldsymbol{a}_{1:T}$，其中每个部分a_t对齐并对应于长度为 T 的驱动视频的帧y_t。然后，我们将音频特征提取应用于 2L 个音频部分$a_{t-L-1:t+L} = \{a_{t-L-1}, \dots, a_t, \dots, a_{t+L}\}$，以帧 t 为中心，获得特征向量$h_t^{(a)}$，其中包含来自过去和未来时间步长的信息。我们采用[16]来提取低级特征，例如 MFCC、信号能量和熵，从而产生特征向量$h_t^{(aL)} \epsilon R^{84}$。然后，我们使用 *DeepSpeech*[18]从每个部分提取位于$a_t{}' \in a_{t-L-1:t+L}$的字符级逻辑。这导致了 2L 个 logits，在连接之后给出了一个特征向量$h_t^{(aH)} \in R^{2L.27}$。在 L = 4 时，我们最终的音频特征向量为$h_t^{(a)} = [h_t^{(aL)^T}; h_t^{(aH)^T}]^T \in R^{300}$。

3.3. HeadGAN 框架

我们基于 GAN 的头部重建系统配备了一个由两种模式驱动的生成器：1)从驱动视频和参考图像中提取的 3D 面部表示，2)可选的来自驱动员的音频特征。给定$x_{t-k:t}$，来自第 t 帧的驱动 3D 人脸表示，与来自过去 k = 2 帧的 3D 人脸按通道连接，参考图像y_{ref}与相应的 3D 人脸表示x_{ref}和音频特征向量$h_t^{(a)}$，我们的生成器产生逼真的图像，给出为

$$\tilde{\mathbf{y}}_t = G(\mathbf{x}_{t-k:t}, \mathbf{y}_{ref}, \mathbf{x}_{ref}, \mathbf{h}_t^{(a)}; \theta_G). \qquad (2)$$

时空体积$x_{t-k:t}$上的条件合成有助于实现跨帧的时间一致性。 参考图像y_{ref}提供了有关源人物的纹理和外观的信息，而音频特征增强了 G 在整个面部（主要是嘴巴区域）的生成能力。 更详细地说，生成器由两个子网络组成：密集流网络 F 和渲染网络 R。有关生成器 G 的概述，请参见图 3。

密集流网络 F。我们的渲染网络 R 依赖于反映源人物外观的高质量视觉特征。尽管如此，我们观察到仅使用编码器从参考图像y_{ref}中提取此类特征并不能很好地利用渲染网络架构的潜力。事实证明，将视觉特征图与所需的头部姿势对齐更有意义，这反映在来自驱动视频的 3D 面部表示x_t中。考虑到这一点，我们提出了一个密集流网络，它学习了一个可用于扭曲视觉特征的流权值。为此，我们通过编码器将参考图像与其对应的 3D 人脸（y_{ref}，x_{ref}）连接起来，以提取三个空间尺度$h_{(1)}$，$h_{(2)}$,$h_{(3)}$的视觉特征图，代表源人物的出现。然后，解码器预测流量w_t，由驱动 3D 面部表示$x_{t-k:t}$引导，该表示通过 SPADE 块[30]注入 F。理想情况下，当应用于参考图像y_{ref}时，这种密集流应该产生源人物的扭曲图像，具有相同的头部姿势和表情，如驱动 3D 人脸表示x_t所示。通过在每个视觉特征图上应用流场，我们得到扭曲的视觉特征$h_t^{(1)}, h_t^{(2)}, h_t^{(3)}$和扭曲的参考图像$y_{ref}$，所有这些取决于第 帧的驱动头部姿势。

渲染网络 R。在我们的 Generator 的核心中，渲染网络旨在将 3D 人脸表示$x_{t-k:t}$转换为源的照片般逼真的图像y_t。这是在高质量音频特征$h_t^{(a)}$和视觉特征图 $h_t^{(1)}$，$h_t^{(2)}$，$h_t^{(3)}$的帮助下实现的。首先，编码器接收$x_{t-k:t}$作为输入并应用一系列具有下采样的卷积层。然后，由交替的 *SPADE*[30]和 *AdaIN*[20]层组成的解码器生成所需的帧$\tilde{y}_t$。这些自适应归一化层能够通过 SPADE 块将 2D 视觉特征图注入渲染网络，并通过 AdaIN 块将 1D 音频特征注入到渲染网络中。与 *SPADE*[30]的原始工作相反，其中所有 SPADE 层的条件输入是相同的分割图，下采样以匹配每层的空间大小，我们利用多个空间尺度的视觉特征图 $h_t^{(1)}$，$h_t^{(2)}$ ，$h_t^{(3)}$，$h_t^{(4)}$，y_t^{ref}作为 SPADE 模块的调制输入。相反，我们将相同的音频特征向量$h_t^{(a)}$传递给所有空间尺度的 AdaIN 块。解码器还配备了用于上采样的 *PixelShuffle layer*[32]，这有助于提高生成样本的质量。

鉴别器 D 和 Dm。图像鉴别器接收合成对（x_t，$\tilde{y}_t$ ）或真实对（x_t，y_t）并学习区分它们。我们使用第二个鉴别器 Dm，它专注于嘴部区域。除了真实的y_t^m或生成的$\tilde{y}_t^m$裁剪的嘴巴区域外，该网络还以音频特征向量$h_t^{(a)}$为条件，该向量在空间上复制，然后按通道连接到裁剪后的图像上。

培训目标。构成生成器的网络 F 和 R 被联合优化。我们通过在扭曲和生成的图像上应用感知和像素损失 L_F^{VGG} 、L_G^{VGG}和L_F^{L1}、L_G^{L1}来训练 *HeadGAN* 重建，如图 3（红色箭头）所示。GAN 铰链损失L_G^{adv}[27]以及特征匹配损失L_G^{FM}[46]进一步增加了结果的真实感。 关于目标函数和网络架构的扩展讨论可以在补充材料中找到。

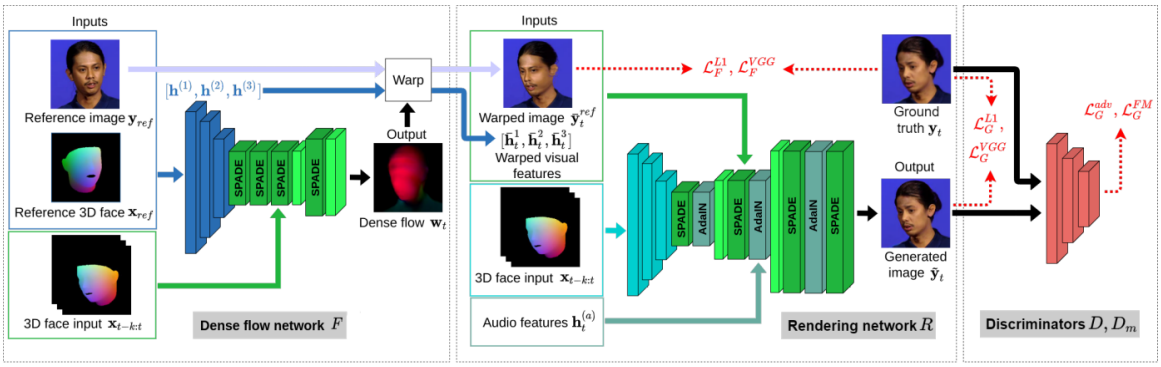

图 3: ***HeadGAN*** 概述。 密集流网络 F 根据 3D 人脸输入，计算用于扭曲参考图像和特征的流场。 然后，渲染网络 R 使用此视觉信息以及音频特征，以便将 3D 人脸输入转换像源照片一般逼真的图像。

3.4. 3D 人脸建模的优势

包含在 3D 面部表示中的语义信息允许我们的密集流网络 F 学习面部区域的精确流，因为它提供了参考图像和驱动图像之间面部点的密集对应关系。 与可以从 3D 网格获得的场景流[28, 26]相比，我们被 F 幻觉的流场存在于缺少 3D 表示的区域，例如头发和上半身，在这些区域中翘曲同样重要。 此外，由于 **3DMM** 允许将人物与表情分开，我们选择在从驱动者提取的 3D 面部上调节渲染网络 R 并适应源的人物特征，使 ***HeadGAN*** 能够解决以下任务中的人物保存问题重建。

最后，使用 **3DMM** 对人脸进行建模使 ***HeadGAN*** 非常适用于视频会议应用程序。 它使“发送者”能够以表达参数$P_t^{exp} \in R^{28}$和相机参数$c_t \in R^7$的形式有效地压缩驱动帧 y_t，总共给出 35 个浮点值。 然后，“接收器”可以使用这些参数来渲染 3D 人脸，并使用生成器来重建驱动框架。单个参考图像需要在会话开始时发送一次。

4. Experiments

4.1. 实现细节

3D 人脸渲染。给定一组相机参数 c 和 3D 面部形状 $\boldsymbol{S} \in \boldsymbol{R^{3N}}$（参见等式 1），我们对 3D 网格进行光栅化并在图像平面中生成可见性蒙版$\boldsymbol{I} \in \boldsymbol{R^{H*W}}$。I 的每个空间位置存储从该像素看到的 3D 面上相应可见三角形的索引。 然后，我们使用 **3DMM** 的平均形状 $\bar{}$x，以找到每个可见三角形中心的归一化 x-y-z 坐标。 通过这种方式，我们获得了一个 3D 人脸表示$\boldsymbol{X} \in \boldsymbol{R^{H*W*3}}$，其中每个像素包含三个坐标。 这些值可以解释为颜色，因此可以解释为 3D 面部的纹理 [52]。

数据集和训练。 我们在 ***VoxCeleb***[29]数据集上训练和评估 ***HeadGAN***，该数据集包含 1,251 个人物的 100,000 多个视频，分辨率为 256 × 256。 我们保持原来的训练和测试分离。 作为预处理步骤，我们为数据集中的每个视频帧计算 3D 人脸图像并提取每帧音频特征向量。 在训练期间，我们执行自我重建，因为我们从目标视频中随机采样参考图像。 这提供了对真实数据的访问，使我们能够设计重建损失项来训练生成器。对于 ***HeadGAN*** 的优化，我们使用 ***ADAM***[24]，其中β_1 = 0.5 ，β_2= 0.999 和学习率 η = 0.0002，来生成器和鉴别器。

4.2. 与基线的比较

重建（自我重建）。 首先，我们将我们的方法与关于自我重建问题的四种最先进的方法进行比较，其中参考人物与驱动人物重合。 在这里，*HeadGAN* 的任务是从 3D 面部表示序列中重建驱动视频，使用单个参考图像访问外观信息。 我们使用 *X2Face*[45]、少镜头 *vid2vid*[43]、双层神经化身[48]和一阶运动模型[34]进行定性和定量比较。 对于[45]和[34]，我们使用作者提供的预训练模型，在 VoxCeleb 上训练。 因为没有可用的检查点，所以我们从头开始训练[43]。 最后，我们使用了[48]的作者提供的模型，在更大的 *VoxCeleb2**[9]上进行了训练。

对于重建的数值评估，我们使用生成的和地面实际帧之间的 L1 距离，以及峰值信噪比(PSNR)和学习的感知图像块相似性(LPIPS)[50]。 我们使用 Frechet 起始距离(FID)[19]和 Frechet 视频距离(FVD)[41]指标访问帧的真实性。 我们使用余弦相似度(CSIM)[11]来衡量人物保留。 结果如表 1 所示，表明 *HeadGAN* 在每个指标中都以显着的幅度优于所有四个基线。

图 4 中所示的示例表明，我们的方法生成的图像比具有更好保留外观特征的基线要真实得多。 我们希冀读者目视检查补充视频中的结果。

重建。重建的目的是将目标序列的头部姿势和面部表情完全转移到源图像中显示的人物上，同时保留后者的人物，因为现在驱动和参考对象不同。为此，我们从 VoxCeleb 测试集中选择了 15 随机（视频、图像）对并进行了重建，每种方法总共生成了大约 6K 帧。除了图像质量(FID)和人物保留(CSIM)外，我们还分别使用平均旋转距离(ARD)和 Action Units Hamming distance(AU-H)[2]进一步评估系统的姿势和表情的可转移性。有关指标的详细信息可以在补充材料中找到。从表 1 中可以看出，*HeadGAN* 在视觉质量方面创造了卓越的样本。 *Bi-layer Neural Avatars*[48]在面部表情转移任务上的表现稍好一些，这可能是因为它是在更大的 *Vox Celeb2*[9]上训练的。然而，它在人物保存方面表现不佳，因为它在面部标志上进行合成，不可避免地传递来自驱动主体的人物相关信息。另一方面，FOMM[34]使用相对关键点位置来解决似乎增加 CSIM 的人物保留问题。尽管如此，这是以姿势转移为代价的，因为模型要求驱动视频第一帧中的人脸与参考人脸具有相同的姿势，这种情况很少见，ARD 也证实了这一点。当 FOMM 使用绝对关键点位置时，为了准确地传递姿势，人物保留问题变得明显，CSIM 下降到 0.587。这种行为可以在图 5 中直观地观察到，其中驱动员的头部几何形状反映在生成的 FOMM-abs 样本中。与基线不同，*HeadGAN* 在成功重建的所有三个要求（姿势、表达转移和人物保存）上都表现良好。这里我们注意到我们省略了与 *MarioNETte*[17]和 *Warp-guided GANs*[14]的比较，因为源代码不公开。

正面化。我们选择使用 3D 面部表示进行条件合成，这使我们可以手动设置所需的头部姿势，而无需驱动框架。通过将相机参数重新设置为正面，我们能够生成参考图像的正面视图。我们将 *HeadGAN* 在正面化任务上与 *pixel2style2pixel*(pSp)[31]和 *Rotate-and-Render*(RaR)[51]进行比较。为此，我们从 *VoxCeleb* 测试分割的每个视频中随机选择一帧并将其正面化。由于此任务没有音频输入，我们训练了 *HeadGAN* 的变体，没有 AdaIN 层的音频特征。然后，我们以度为单位，使用 FID 测量生成的样本的照片真实度、使用 CSIM 的人物保留和平均旋转误差 (ARE)作为与正面姿势的偏差。我们将正面姿态定义为来自 3DMM 拟合的相机旋转参数中的零欧拉角。因此，ARE 在此仅用作参考，作为健全性检查。结果显示在表 2 中。 *HeadGAN* 在 CSIM 和 RaR 上的表现同样出色，并且在图像质量和正面化精度方面超过了基线。请参见图 6 进行视觉比较。

	Reconstruction						Reenactment			
Method	L1 ↓	PSNR ↑	LPIPS ↓	FID ↓	FVD ↓	CSIM ↑	FID ↓	CSIM ↑	ARD ↓	AU-H ↓
X2Face [45]	13.49	20.69	0.260	130.2	697	0.600	122.1	0.520	4.39	0.346
fs-vid2vid [43]	17.15	18.52	0.197	62.8	471	0.542	-	-	-	-
*Bi-layer** [48]	*12.18*	20.19	*0.152*	92.2	394	0.590	172.8	0.563	**1.01**	**0.296**
FOMM [34]	12.34	*20.93*	0.153	*64.9*	*338*	*0.754*	*63.7*	**0.765**	12.53	0.400
HeadGAN	**11.32**	**21.46**	**0.112**	**36.1**	**254**	**0.807**	**58.0**	*0.688*	*1.35*	*0.326*

表 1：VoxCeleb [29] 测试集重建和重演任务的定量结果。

图 4：在重建（自我重演）任务上与基线的定性比较。

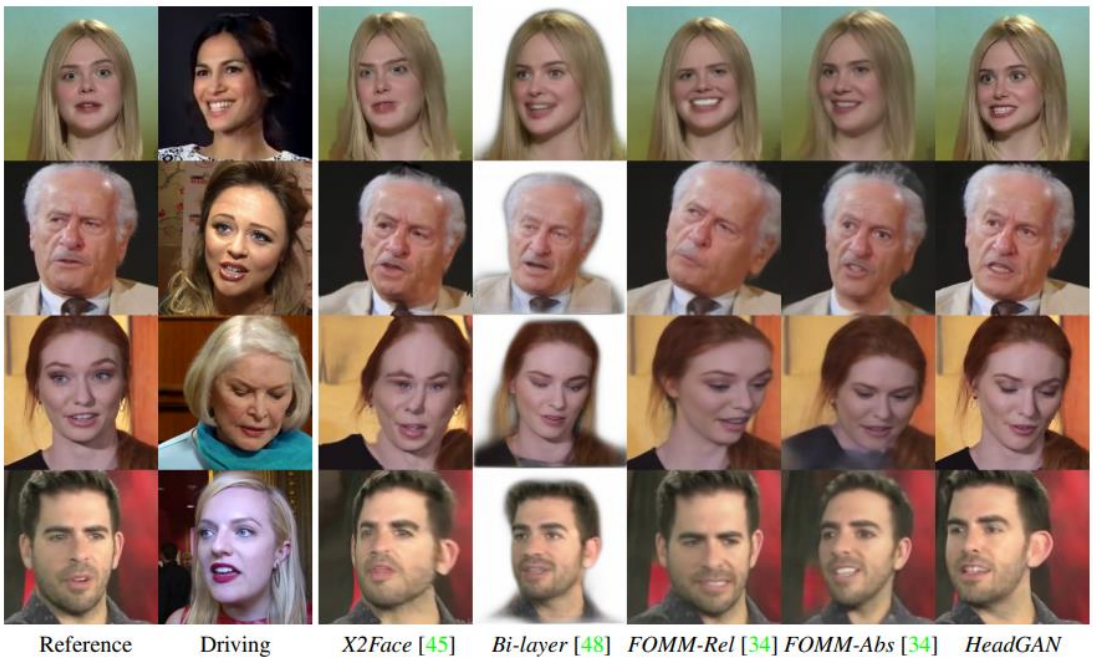

图 5：在重演任务上与基线的定性比较。 FOMM-Rel 指相对关键点坐标策略，FOMM-Abs 指绝对关键点坐标，用于一阶运动模型 [34]。

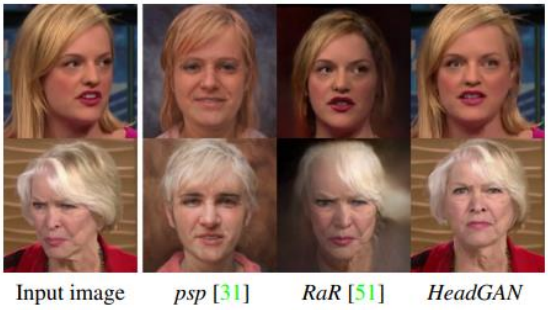

图 6：正面化的定性比较。

Method	FID ↓	CSIM ↑	ARE ↓
psp [31]	147.8	0.130	2.66
RaR [51]	*88.4*	*0.753*	*2.65*
HeadGAN	**30.1**	**0.766**	**0.76**

表 2：正面化的定量结果。

4.3. 图像表达和姿势编辑

我们的模型可以进一步用作图像编辑工具。 给定源图像y_{ref}及其形状和相机参数y_{ref}^{id}、p_{ref}^{exp}和c_{ref}，首先我们渲染相应的 3D 人脸表示x_{ref}。 然后，我们手动重新调整表情或相机参数并渲染一个伪驱动 3D 人脸x_t。 我们通过生成器传递x_t，y_{ref}，x_{ref}以获得具有新颖表情或姿势的合成图像。 在图 7 中，我们显示了在处理表情的前三个主成分(p. c.)和摄像机角度后的结果。

我们进行了消融研究，以评估 1)密集流网络 F 的重要性，2)3D 面部表示与 2D landmark 草图相比的优势，即[43]和[48]使用的输入， 3）音频模态的贡献。为了评估 F 的重要性，我们移除了它的解码层。我们保留了它的编码器，它从参考图像中提取外观特征图。然后，我们没有扭曲这些特征图和参考图像，而是将它们直接传递到渲染网络 R 的 SPADE 层。此外，我们通过对草图图像进行条件合成，通过连接 2D 绘制，实现了具有 landmark 的 *HeadGAN* 变体带边缘的 landmark[43]。从表 3 中可以看出，完整模型优于所有变体。在分数方面，我们观察到流网络 F 是我们系统的重要组成部分。在图 8b 中，使用 3D 人脸表示（而不是 landmark）减轻了人物保存问题，并且可以在视觉上更好地理解。

唇读实验。我们通过使用外部唇读网络对合成视频进行分类，进一步定量评估音频输入对我们系统的贡献。为此，我们选择了 BBC 数据集[10]的 25 个词类，并在默认训练分组上训练了唇读分类器[36]。 之后，我们使用视频中的随机帧作为参考，重建了 BBC 数据集的测试拆分。我们报告了真实测试样本的唇读准确率为 97%，使用完整模型生成的样本为 82%，在不考虑音频输入的情况下由变异产生的合成数据为 73%。 这些结果表明，音频模式在很大程度上有助于产生更合理的嘴唇运动。

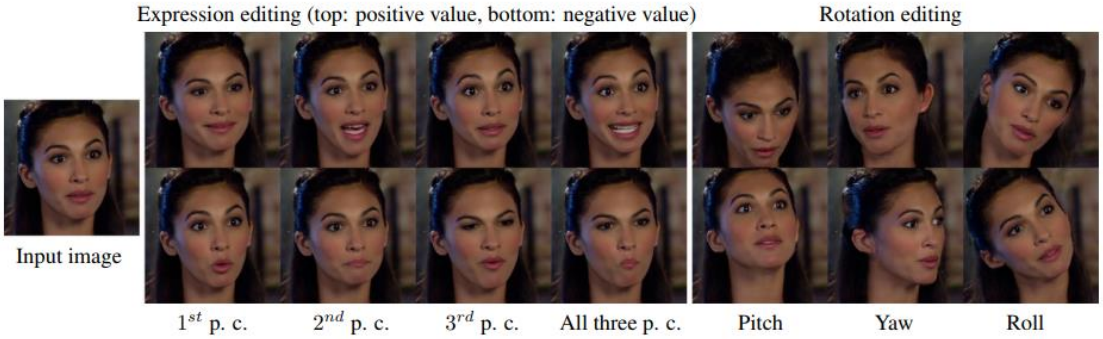

图 7：表情和姿势编辑。 请注意，我们使用我们模型的无音频变体 (AdaIN) 来完成此任务。

Method	FID ↓	FVD ↓	CSIM ↑
HeadGAN w/o network F	63.3	473	0.307
HeadGAN w/ landmarks	55.7	371	0.699
HeadGAN w/o audio input	55.1	356	0.687
HeadGAN	**50.9**	**334**	**0.716**

表 3：消融研究数值结果。 请注意我们在这个实验中训练了半个 epochs 的模型。

(a) Significance of dense flow network F in image quality.

(b) The identity preservation problem becomes prominent when conditioning on facial landmarks, instead of the 3D face representation.

图 8：***HeadGAN*** 组件的重要性。

5. Conclusion

我们展示了 ***HeadGAN***，这是一种新颖的一次性头部动画方法，由 3D 面部数据和音频特征驱动。 与 SOTA 方法相比，我们的框架表现出卓越的再现性能和更高的真实感。 我们的方法可以进一步用于重建、姿势和面部表情编辑以及正面化。

References

……

翻译：任辽

2021/11/12