E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2024)09-2494-19

论文引用格式:Gao X, Liu D Y and Zhang J Y. 2024. Multi-modal digital human modeling, synthesis, and driving: a survey. Journal of Image and Graphics, 29(09):2494-2512(高玄, 刘东宇, 张举勇. 2024. 多模态数字人建模、合成与驱动综述. 中国图象图形学报, 29(09):2494-2512)
[DOI:10.11834/jig.230649]

多模态数字人建模、合成与驱动综述

高玄,刘东宇,张举勇*

中国科学技术大学安徽省图形计算与感知交互重点实验室,合肥　230026

摘　要:多模态数字人是指具备多模态认知与交互能力,且有类人的思维和行为逻辑的真实自然虚拟人。近年来随着计算机视觉与自然语言处理等领域的交叉融合以及蓬勃发展,相关技术取得显著进步。本文讨论在图形学和视觉领域比较重要的多模态人头动画、多模态人体动画以及多模态数字人形象构建3个主题,介绍其方法论和代表工作。在多模态人头动画主题下介绍语音驱动人头和表情驱动人头两个问题的相关工作。在多模态人体动画主题下介绍基于循环神经网络(recurrent neural networks,RNN)的、基于Transformer的和基于降噪扩散模型的人体动画生成。在多模态数字人形象构建主题下介绍视觉语言相似性引导的虚拟形象构建、基于多模态降噪扩散模型引导的虚拟形象构建以及三维多模态虚拟人生成模型。本文将相关方向的代表性工作进行介绍和归类,对已有方法进行总结,并展望未来可能的研究方向。

关键词:虚拟数字人建模;多模态角色动画;多模态生成与编辑;神经渲染;生成模型;神经隐式表示

Multi-modal digital human modeling, synthesis, and driving: a survey

Gao Xuan, Liu Dongyu, Zhang Juyong*

Key Laboratory of Computer Graphics and Perception Interaction in Anhui Province, University of Science and Technology of China, Hefei　230026, China

Abstract: A multimodal digital human refers to a digital avatar that can perform multimodal cognition and interaction and should be able to think and behave like a human being. Substantial progress has been made in related technologies due to cross-fertilization and vibrant development in various fields, such as computer vision and natural language processing. This article discusses three major themes in the areas of computer graphics and computer vision: multimodal head animation, multimodal body animation, and multimodal portrait creation. The methodologies and representative works in these areas are also introduced. Under the theme of multimodal head animation, this work presents the research on speech- and expression-driven head models. Under the theme of multimodal body animation, the paper explores techniques involving recurrent neural network (RNN)-, Transformer-, and denoising diffusion probabilistic model (DDPM)-based body animation. The discussion of multimodal portrait creation covers portrait creation guided by visual-linguistic similarity, portrait creation guided by multimodal denoising diffusion model, and three-dimensional (3D) multimodal generative models on digital portraits. Further, this article provides an overview and classification of representative works in these research directions, summarizes existing methods, and points out potential future research directions. This article delves into key directions in the field of multimodal digital humans and covers multimodal head animation, multimodal body animation, and the

收稿日期:2023-09-12;**修回日期**:2024-03-15;**预印本日期**:2024-03-22

*通信作者:张举勇　juyong@ustc.edu.cn

基金项目:国家自然科学基金项目(62122071,62272433)

Supported by:National Natural Science Foundation of China (62122071, 62272433)

construction of multimodal digital human representations. In the realm of multimodal head animation, we extensively explore two major tasks: expression- and speech-driven animation. For explicit and implicit parameterized models for expression-driven head animation, mesh surfaces and neural radiance fields (NeRF) are used to improve the rendering effects. Explicit models employ 3D morphable and linear models but encounter challenges, such as weak expressive capacity, nondifferentiable rendering, and difficult modeling of personalized features. By contrast, implicit models, especially those based on NeRF, demonstrate superior expressive capacity and realism. In the domain of speech-driven head animation, we review 2D and 3D methods, with a particular focus on the important advantages of NeRF technology in enhancing realism. 2D speech-driven head video generation utilizes techniques, such as generative adversarial networks and image transfer, but depends on 3D prior knowledge and structural characteristics. On the other hand, methods using NeRF, such as audio driven NeRF for talking head synthesis (AD-NeRF) and semantic-aware implicit neural audio-driven video portrait generation (SSP-NeRF), achieve end-to-end training with differentiable NeRF. This condition substantially improves rendering realism while still addressing challenges associated with slow training and inference speeds. Multimodal body animation focuses on speech-driven body animation, music-driven dance, and text-driven body animation. We focus on the importance of learning speech semantics and melody and discuss the applications of RNN, Transformer, and denoising diffusion models in this field. Transformer gradually replaces RNN as the mainstream model, which gains notable advantages in sequence signal learning through attention mechanisms. We also highlight the body animation generation based on denoising diffusion models, such as free-form language-based motion synthesis and editing (FLAME), motion diffusion model (MDM), and text-driven human motion generation with diffusion model (MotionDiffuse), and multimodal denoising networks under music and text conditions. In the realm of the construction of multimodal digital human representations, the article emphasizes virtual-image construction guided by visual-language similarity and denoising of diffusion models. In addition, the demand for large-scale, diverse datasets in digital human representation construction is addressed to foster powerful and universal generative models. The three key aspects of multimodal digital humans are systematically explored: head animation, body animation, and digital human representation construction. In summary, explicit head models, although simple, editable, and computationally efficient, lack expressive capacity, and face challenges in rendering, especially in modeling facial personalization and nonfacial regions. By contrast, implicit models, especially those using NeRF, demonstrate stronger modeling capabilities and realistic rendering effects. In the realm of speech-driven animation, NeRF-based solutions for head animation overcome the limitations of 2D speaker and 3D digital head animation and achieve more natural and realistic speaker videos. Regarding body animation models, Transformer gradually replaces RNN, whereas denoising diffusion models can be used to potentially address mapping challenges in multimodal body animation. Finally, digital human representation construction faces challenges, with visual-language similarity and denoising diffusion model guidance showing promising results. However, the difficulty lies in the direct construction of 3D multimodal virtual humans due to the lack of sufficient 3D virtual human datasets. This study comprehensively analyzes various issues and provides clear directions and challenges for future research. In conclusion, should focus on future developments in multimodal digital humans. Key directions include improvement of 3D modeling and real-time rendering accuracy, integration of speech-driven and facial expression synthesis, construction of large and diverse datasets, exploration of multimodal information fusion and cross-modal learning, and addressing ethical and social impacts. Implicit representation methods, such as neural volume rendering, are crucial for improved 3D modeling. Simultaneously, the construction of larger datasets poses a formidable challenge in the development of robust and universal generative models. Exploration of multimodal information fusion and cross-modal learning allows models to learn from diverse data sources and present a range of behaviors and expressions. Attention to ethical and social impacts, including digital identity and privacy, is crucial. Such research directions should serve as guide the field toward a comprehensive, realistic, and universal future, with profound influence on interactions in virtual spaces.

Key words: virtual human modeling; multimodal character animation; multimodal generation and editing; neural rendering; generative models; neural implicit representation

0 引 言

数字虚拟人在增强现实/虚拟现实(augmented reality/virtual reality, AR/VR)、游戏和影视制作等场景中有着广泛应用,是图形学、视觉和人机交互等领域的经典研究问题。多模态虚拟人研究如何让虚拟人具备多模态认知与交互能力,同时让虚拟人表现出类人的思维和行为逻辑,有着广阔的应用前景。多模态数字人研究中的核心问题是找出模态信号与数字人形象/动画的映射关系并保证数字人的真实感。最近几年由于计算机视觉、自然语言处理相关领域的快速发展,多模态数字人的相关研究也有了巨大进步,如图1所示。具体来说,多模态数字人是指具备多模态认知与交互能力,且有类人的思维和行为逻辑的真实自然虚拟人。

图1 多模态数字人的相关研究

Fig. 1 Research on multimodal digital humans

多模态数字人的研究主要聚焦于人头动画、人体动画以及形象构建这3类任务上。本文第1节介绍人头动画内容,第2节介绍人体动画内容,第3节介绍形象构建内容。

多模态人头动画主要包含两类任务,一类是表情驱动人头;一类是语音驱动人头。由于人头表示的不同,多模态人头动画的技术路线也有很大的区别。表情驱动人头的核心就是构建人头参数化模型,根据模型表示的不同可以划分为显式参数化人头模型与隐式参数化人头模型两类。语音驱动人头的技术路线根据虚拟人头表示主要划分为3种,分别是语音驱动二维说话人视频生成、语音驱动三维数字人头几何动画和语音驱动人头神经辐射场3类。

多模态人体动画本质上是序列生成任务,相关方法也主要依序列生成模型结构的不同进行区分。早期的主流方法一般使用循环神经网络(recurrent neural networks, RNN)(Medsker和Jain, 2001)模型建模这一问题。Transformer(Vaswani等, 2017)相比于RNN模型,扩展性更好、对长程信息以及位置信息的捕捉更鲁棒,逐渐取代了RNN模型。近年来出现的降噪扩散模型(Ho等, 2020)具有极强的分布建模的能力,非常适合解决多模态人体动画中一到多的映射难题,最近的一些工作也展示出其在多模态人体动画任务中的潜力。本文首先介绍语音、音乐和文本信号的学习与编码,然后按照模型出现时间顺序介绍基于RNN的运动生成、基于Transformer的运动生成以及基于降噪扩散模型的运动生成。

多模态数字人形象构建这一问题是随着最近的视觉语言模型的巨大突破而出现的一类新问题,主要技术路线有两种,一种是使用视觉语言相似性引导虚拟人形象构建,主要依赖CLIP(contrastive language-image pre-training)(Radford等, 2021)空间下相似性损失函数的引导。另一种基于多模态降噪扩散模型引导的虚拟形象构建,通过SDS(score distillation sampling)(Poole等, 2022)来引导形象生成。也有工作选择直接构建多模态的三维虚拟人生成模型来解决这一问题。本文首先讨论使用视觉语言相似性引导虚拟人形象构建,其次讨论基于多模态降噪扩散模型引导的虚拟形象构建,最后讨论直接构建多模态的三维虚拟人生成模型的工作。

图2给出了多模态数字人的研究发展图,主要展示相关的代表性工作。早期研究(Fisher, 1968)提出了一种通过构建音素(phoneme)与视素(viseme)的对应关系来描述语音与嘴型的联系。然而,这种简单的对应方式存在局限性,无法捕捉不同音素组合带来的细微嘴部运动变化。随后,Blanz和Vetter(1999)提出了三维可变形模型(3D morphable model, 3DMM),这成为三维人脸重建领域最常用的参数化人脸模型。另一项工作JALI(jaw and lip)(Edwards等, 2016)建立了语音与三维人脸FACS(facial action coding system)基表示之间的映射关系,通过音素、音量、音高和共振峰等信息的强制对齐,得到了FACS人脸模型的激活单元(action unit)。这种模型构建过程需要大量动画师的人工操作,制作成本较高。

过去采用的人体动画生成模型多基于RNN结构。然而,Transformer(Vaswani等, 2017)作为一种

新型的序列信号学习模型逐渐取代了RNN模型。本文详细介绍了动作条件下使用Transformer解决人体动画生成问题的经典工作ACTOR(action-conditioned 3d human motion synthesis with Transformer VAE)(Petrovich等,2021)等。

对人头动画的一些最新研究也值得关注。Jamaludin等人(2019)首次采用卷积神经网络(convolutional neural network,CNN)编码语音特征与身份特征来驱动图像生成。而NerFACE(Gafni等,2021)是最早提出神经辐射场(neural radiance fields,NeRF)人头模型的研究之一,通过表情系数与位置编码拼接,利用多层感知器(multilayer perceptron, MLP)映射为RGB和体密度,从表情系数序列生成真实的人头动态视频。进一步地,AD-NeRF(audio-driven NeRF)(Guo等,2021b)首次将神经辐射场技术应用到语音驱动三维数字人的任务中。

在多模态数字人形象构建方面,StyleGAN-NADA(Gal等,2021)提出了如何利用CLIP(contrastive language-image pre-training)(Radford等,2021)相似性顺势函数微调StyleGAN,而降噪扩散模型也广泛应用到人体运动合成领域。近期的研究T2M-GPT(Zhang等,2023a)使用VQ-VAE(vector quantized variational autoencoder)(van den Oord等,2017)构建了人体运动codebook,并以CLIP编码的文本特征为条件,自回归生成人体code序列。最后,Instruct-NeRF2NeRF(Haque等,2023)利用Instruct-Pix2Pix(Brooks等,2023)引导三维人像和场景生成,实现了通过命令语句进行三维编辑。

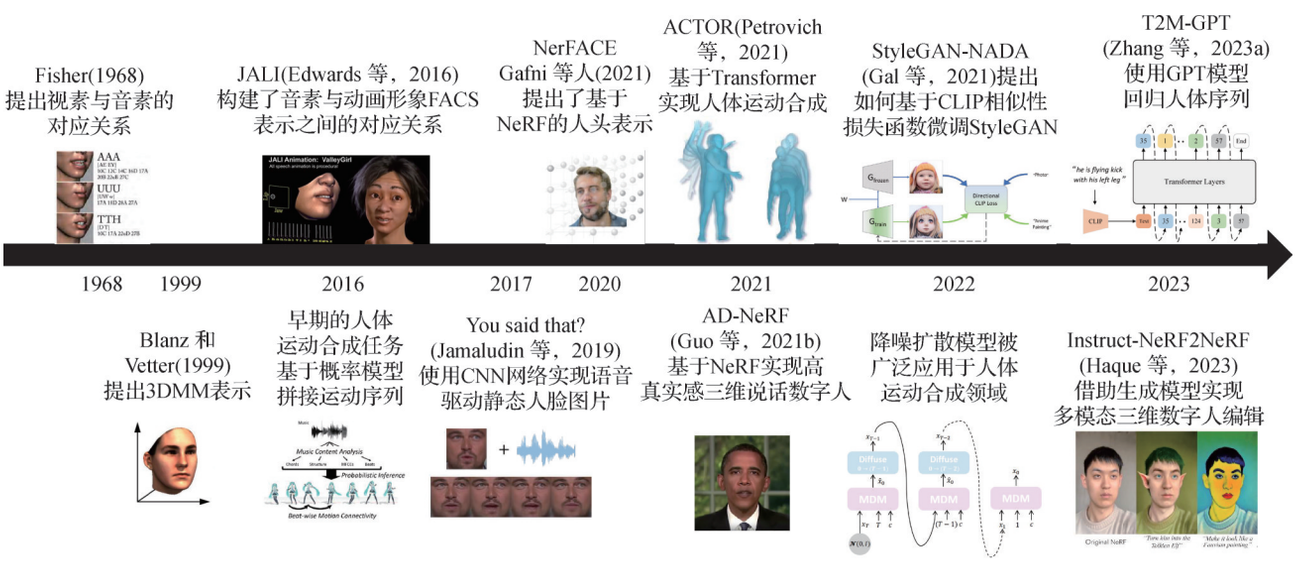


图2　多模态数字人建模、合成与驱动发展

Fig. 2　Development of multi-modal digital human modeling, synthesis and driving

1　多模态人头动画

多模态人头动画主要包含两类任务,一类是使用表情系数驱动人头;另一类是使用语音驱动人头。表情系数和语音可看做驱动人头的两种模态。另外,有的工作中将图像的身份信息和反射率、光流等信息作为多模态的输入。

1.1　表情驱动人头

表情驱动人头是指通过表情系数来驱动虚拟人头做出相应表情动画,这一任务的核心是构建参数化人头模型,参数化人头模型的构建主要有显式和隐式两种方案。显式表示用网格曲面(mesh)来表示人头,有线性模型和非线性模型两种思路;隐式表示一般用NeRF、SDF(signed distance fields)或者Occupancy field表示人头。

1.1.1　显式参数化人头模型

自Blanz和Vetter(1999)提出三维可变形模型(3DMM)以来,这一表示已成为三维人脸重建中最常用的参数化人脸模型。3DMM方法首先用激光扫描仪采集几百个人在中性表情下的几何和反照率信息,然后对这些数据进行主成分分析(principle component analysis, PCA),从而将人脸几何和反照率信息表示为

$$\begin{cases} p = \bar{p} + A_{\mathrm{id}}\alpha_{\mathrm{id}} \\ b = \bar{b} + A_{\mathrm{alb}}\alpha_{\mathrm{alb}} \end{cases} \tag{1}$$

式中，p表示人脸几何，$\bar{p}$表示中性脸几何，A_{id}表示几何身份基，α_{id}表示身份系数；b表示人脸反照率，$\bar{b}$表示中性反照率，A_{alb}表示反照率基，α_{alb}表示反照率系数。

3DMM表示的一个缺陷是其只能表示中性脸，不能编码人脸的表情信息。Chu等人(2014)将3DMM扩展到将表情也用PCA表示，具体为

$$p = \bar{p} + \begin{bmatrix} A_{\mathrm{id}} & A_{\mathrm{exp}} \end{bmatrix} \begin{bmatrix} \alpha_{\mathrm{id}} \\ \alpha_{\mathrm{exp}} \end{bmatrix} \tag{2}$$

该表示中身份和表情是线性相加的关系，因此身份和表情不能有效分离。Cao等人(2014)将身份和表情对几何的影响用双线性模型来表示。为此，作者用Kinect采集并重建了150个人的47种表情对应的三维人脸几何数据，之后对由这些数据组成的三维张量(身份、表情和顶点三个维度)进行高维奇异值分解(*n*-mode singular value decomposition)，得到主张量A_r。利用主张量，三维人脸几何可以表示为

$$p = A_r \times w_{\mathrm{id}} \times w_{\mathrm{exp}} \tag{3}$$

式中，w_{id}和w_{exp}分别表示身份和表情系数，用来控制三维人脸几何的变化。

也有工作指出这样的模型的表达能力受到线性关系的限制，提出了非线性的人脸模型表示(Ranjan等，2018；Tran和Liu，2018；Guo等，2021a)，如图3所示，非线性3DMM用神经网络建模几何和纹理信息，相比于线性模型表达能力更强。

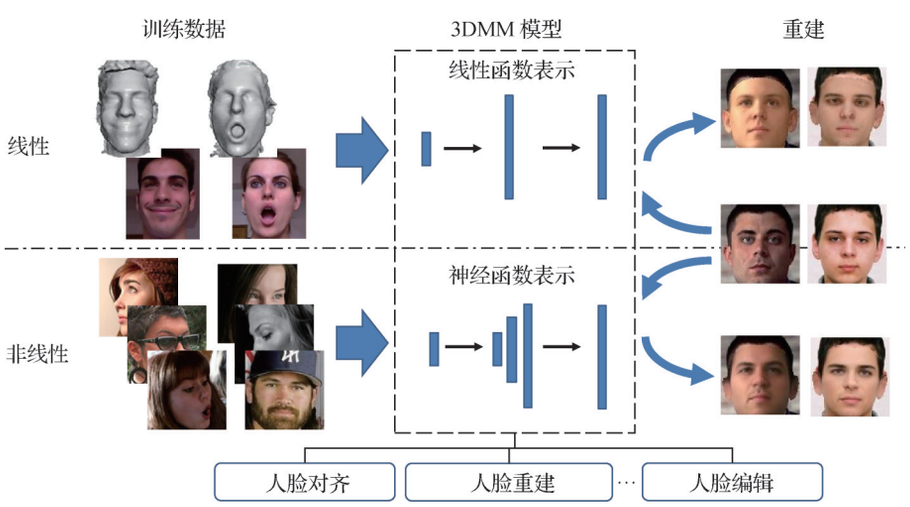


图3　非线性3DMM与线性3DMM的比较

Fig. 3　Comparison of non-linear 3DMM and linear 3DMM

FLAME(Li等，2017)通过LBS(linear blend skinning)表示脖子和颚的铰接变形，使表情动画更加自然。其计算式可以写为

$$M(\beta,\theta,\psi) = H\left(T_P(\beta,\theta,\psi), J(\beta), W\right) \tag{4}$$

$$T_P(\beta,\theta,\psi) = \bar{T} + B_S(\beta,S) + B_P(\theta,P) + B_E(\psi,E) \tag{5}$$

式中，β表示身份系数，θ表示颚、头和眼睛的角度信息，ψ表示表情系数。H表示蒙皮计算，W表示蒙皮权重。$J(\beta)$表示关节点位置，其仅与β有关。$B_S(\beta,S), B_P(\theta,P), B_E(\psi,E)$分别是基于身份系数、角度信息以及表情系数计算出的基准几何校正项。

显式模型使用简单、可编辑性强且计算高效，但是模型表达能力弱，渲染不具有可微性，且在建模人脸个性化特征以及表示非人脸区域(如头发等)上困难比较多。

1.1.2　隐式参数化人头模型

为了解决显式网格人头模型在表达能力上的不足，越来越多的研究者采用隐式表示来处理人头参数化的问题。i3DMM(Yenamandra等，2021)是第1个隐式的三维可变形人头模型，基于SDF表征人头几何。IM Avatar(Zheng等，2022)提出了一种隐式的LBS模型，更为真实地建模了人脸的几何纹理细节。同时也极大提高了表情外插时的鲁棒性。

NeRF(Mildenhall等，2020)提出后，由于NeRF

渲染真实、完全可微的性质，受到越来越多研究人员的青睐。这类方法一般都是学习如下映射，具体为

$$D_\theta(p,v,w)=(c,\sigma) \tag{6}$$

式中，p 表示采样点位置，v 表示光线入射方向，w 表示表情系数，c 和 σ 表示颜色和体密度，用于体渲染。

NerFACE（Gafni 等，2021）是最早的 NeRF 人头模型，通过将表情系数与位置编码进行拼接，通过 MLP 映射为 RGB 和体密度，输入表情系数序列，通过体渲染得到了真实的人头动态视频。

HeadNeRF（Hong 等，2022b）提出了泛化的人头 NeRF 表示，实现了人头身份、表情、纹理、光照属性的控制，并且实现了 NeRF 人头的实时渲染。如图 4 所示，HeadNeRF 先根据身份、表情、纹理和光照体渲染低分辨率特征图 $\boldsymbol{I}_F$，然后通过超分模块 π_ψ 将 $\boldsymbol{I}_F$ 转换为 RGB 图像。

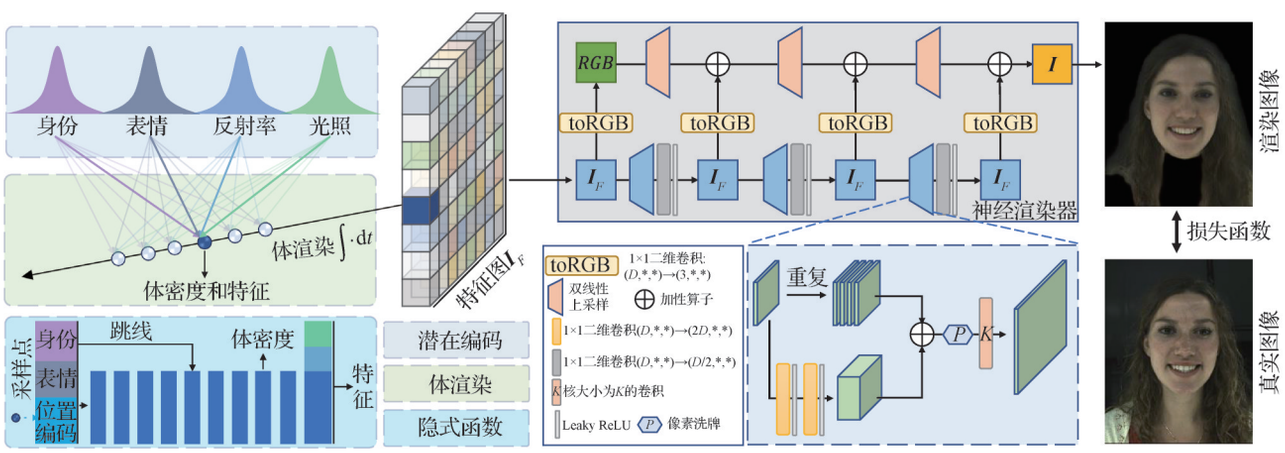


图 4　HeadNeRF 流程图

Fig. 4　HeadNeRF pipeline diagram

Next3D（Sun 等，2023）采用显隐式结合的策略，借助 FLAME 模型身份表情的解耦性，实现了更具真实感的 NeRF 人头参数化模型。NeRFBlendShape（Gao 等，2022）通过构建多分辨率哈希基底，将传统的 blendshape 模型推广到 NeRF 表示上，同时把 NeRF 人头模型的构建速度从原来的几小时缩短到了 10～20 min。

隐式表示的人头模型相比于显式模型渲染更加真实，建模能力更强，但是由于显式信息的缺失，如何直接编辑人头几何与纹理仍是一个尚未完全解决的问题。

1.2　语音驱动人头

基于语音的化身驱动是指输入一段语音，生成某人说话视频的技术。目前，主流的语音驱动人头的方案有 3 种，第 1 种是语音驱动二维说话人视频生成；第 2 种是语音驱动三维数字人头几何动画；第 3 种是语音驱动人头神经辐射场。

1.2.1　语音驱动二维说话人视频生成

二维的说话人视频生成工作经常使用生成对抗网络（Goodfellow 等，2014）或图像迁移（Isola 等，2017）作为其核心技术。You said that?（Jamaludin 等，2019）最早采用卷积神经网络编码语音特征与身份特征来驱动图像，如图 5 所示，输入语音的 MFCC（Mel-frequency cepstral coefficients）特征以及驱动身份的一幅图像，通过卷积神经网络进行编解码，得到一个人的说话视频。

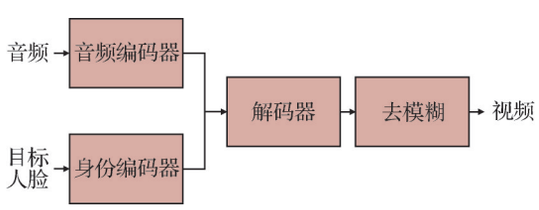


图 5　“You said that?”算法

Fig. 5　The algorithm pipeline of “You said that?”

Chen 等人（2019）和 Zhou 等人（2020）采用关键点作为中间表示来实现说话人视频生成。Wav2Lip（Prajwal 等，2020）采用对比学习的策略，预训练了语音视觉同步网络 SyncNet（Chung 和 Zisserman，2016），然后使用这一专家网络来隐式地保证语音和嘴型的一致性。也有一些工作着眼于小样本的二维说话人构建。如 Zakharov 等人（2019）采用了元学习

预训练高容量的生成器和判别器，然后微调的策略；Wang等人（2021b）通过预测光流实现了单帧的说话人生成。最近随着扩散模型（Ho等，2020）的发展，也有基于扩散模型的二维说话人生成方法（Shen等，2023）。通过将语音信号和人脸关键点信号输入到降噪扩散模型的降噪网络中，对视频进行符合语音的嘴部预测。

除了提高二维说话人视频生成的质量，如何实现可控的二维语音驱动人也是一个研究热点问题。Ji等人（2021）实现了利用情绪标签编辑二维说话人，而Zhou等人（2021）实现了说话人的头部姿态控制。

二维方法在生成说话人视频任务上速度快、计算资源需求低、易于实现，但它们也存在一些限制。这些方法受制于缺乏三维先验知识和卷积神经网络结构特性，在细节和真实感方面表现不尽人意。具体来说，二维说话人生成模型的局限性在于缺乏对三维人脸几何的理解，导致生成的人像缺乏三维逼真度。此外，模型容量和数据集限制使得生成结果难以达到高分辨率。另外，卷积神经网络结构的缺陷也会导致二维模型出现“texture sticking”问题，即人脸细节似乎与图像的绝对坐标粘连在一起，而非真实地呈现在人脸几何表面上。这些因素共同影响了二维方法在生成逼真说话人视频方面的表现。

1.2.2 语音驱动三维数字人头几何动画

早期人们通过构建音素（phoneme）与视素（viseme）的对应关系（Fisher，1968）来描述语音与嘴型的对应关系。这样的对应太过简单，缺少了不同音素组合带来的对嘴型细微变化的建模。JALI（Edwards等，2016）构建了语音与三维人脸激活编码系统的激活单元的映射关系，根据强制对齐的音素、音量、音高和共振峰等信息，映射得到人脸激活编码系统的激活单元（action unit），同时用户可以控制下巴和嘴唇的运动幅度，以此来建模不同的说话风格，这样的模型构建需要动画师的大量人工操作，制作成本很高。

随着深度学习技术的发展，基于数据学习语音与人脸三维变形的关系渐渐成为科研领域的主流思路。VOCA（voice operated character animation）（Cudeiro等，2019）提出使用DeepSpeech（Hannun等，2014）来编码语音，一方面可以消除语音中的身份信息，实现较好的跨语言跨身份驱动；另一方面可以降低网络的学习压力，提升预测质量。VOCA使用高质量扫描数据进行监督，输入语音特征，通过1D CNN融合时间窗内的语音特征，然后用一个小规模的MLP进行解码，取得了较好的语音驱动效果。Neural Voice Puppetry（Thies等，2020）在VOCA的管线基础上进一步完善，采用两阶段的训练思路，先以RGB视频跟踪的参数化网格以及3DMM系数为监督，训练语音到参数化网格的映射网络，然后对网格做投影渲染，采用神经渲染生成真实人脸，实现了基于RGB视频的三维数字人形象建模。Synthesizing Obama（van den Oord等，2017）使用了一套4阶段的管线用十几个小时的Obama演讲数据训练了语音驱动Obama的个性化模型，先用RNN基于语音预测嘴部关键点，然后基于嘴部关键点合成嘴部纹理，再根据语音内容重新对齐了时间域，最后进行人像组合生成最终视频。

随着Transformer（Vaswani等，2017）在自然语言处理领域的发展，其架构也应利用到语音驱动三维数字人这一问题上，FaceFormer（Fan等，2021）先采用自注意力机制对语音进行编码，通过手动设定的attention map来指定生成信号与驱动信号之间的时域对应关系，然后利用其设计的多模态注意力机制根据语音编码的结果对人脸动画进行自回归的解码。CodeTalker（Xing等，2023）在这一框架下进一步改进，利用VQ-VAE（van den Oord等，2017）离散化的思想，将人脸动画序列编码为有限维的字典，一定程度上解决了语音驱动问题多到多映射学习的困难，提升了生成结果的质量。

MeshTalk（Richard等，2021）注意到人脸动画存在语音强相关的部分，也有语音弱相关的部分（如眨眼等），所以提出了先构建人脸动画类别隐空间（categorical space），然后采用类PixelCNN（van den Oord等，2016）的策略在这一隐空间上进行自回归生成，获得了更真实、更具个性化的生成结果。

现有工作已经可以实现比较真实的语音驱动三维数字人几何结果，但这一类方法仍面临一些困难，一是这一类方法依赖三维人脸几何监督，需要足够高精度的重建技术支持；二是由于参数化网格模型表达能力的限制，三维几何动画序列到真实视频的转化仍然有较多困难。

1.2.3 语音驱动人头神经辐射场

相比于之前的语音驱动三维数字人工作，基于神经辐射场的语音驱动人头具有以下特点：1）由于

NeRF 的可微性，可以直接使用 RGB 图像监督进行端到端的训练，不会有因重建而带来的精度损失；2）由于 NeRF 在场景建模的高表达能力，渲染结果足够真实，且有良好的三维一致性。这类方法一般可以表示为

$$D_\theta(p, v, a) = (c, \sigma) \tag{7}$$

式中，p 表示采样点位置，v 表示光线入射方向，a 表示语音特征，c 和 σ 表示颜色和体密度，用于体渲染。

AD-NeRF（Guo 等，2021b）最早将神经辐射场技术应用到语音驱动三维数字人任务中，如图 6 所示，基于语音预测动态神经辐射场表示下的人头和身体，然后将渲染结果进行拼接。具体来说，AD-NeRF 将语音 DeepSpeech 特征与位置编码一起输入给隐式函数，预测 RGB 信息与体密度，实现了高真实感的三维说话数字人建模。SSP-NeRF（Liu 等，2022）在 AD-NeRF 的基础上增加了语义分割图渲染分支（parsing branch），然后增加人脸语义图进行监督训练。使得模型可以更好地建模人脸细节语义。DFRF（Shen 等，2022）解决了语音驱动 NeRF 人头的小样本训练问题，提出可以先预训练一个基本说话人模型，然后用几十秒的小样本对模型进行微调，以此构建可语音驱动的动态人头神经辐射场。

GeneFace（Ye 等，2023）提出了一种三阶段的模型构建思路，先训练泛化的语音驱动变分运动生成器（variational motion generator），根据语音信号生成关键点序列，然后对关键点序列做特定身份的域迁移，最后以关键点为条件控制动态人头神经辐射场，渲染得到说话人视频序列。这一生成模型可以更好地解决语音驱动人的多到多映射的困难，取得了更好的渲染结果。

RAD-NeRF（Tang 等，2022）将高维的语音驱动人像表示分解为低维的可学习体素场，利用高效的采样和体渲染实现，达到实时渲染的目的，一定程度上解决了语音驱动 NeRF 推理速度慢的问题。

DialogueNeRF（Zhou 等，2022）建模了对话场景下听者对讲者做出反应的过程，通过提取讲者的视觉和语音信号，以此引导听者人头 NeRF 的变化。

语音驱动人头神经辐射场的路线相比于前两种渲染真实感更高，但是训练和推理速度较慢，进一步提升语音驱动人头神经辐射场的训练和推理效率也是这一领域的热点问题。

表 1 对部分人头驱动方法进行了比较，使用峰值信噪比（peak signal-to-noise ratio，PSNR）指标对相应方法进行评估，概括了 2D、3D 和 NeRF 等方法的优缺点。

2 多模态人体动画

目前研究较多的多模态人体动画生成问题主要有：1）语音生成人体动画，使人体动画与语音自然协同；2）音乐生成舞蹈，使舞蹈符合音乐的风格与节拍；3）文本生成人体动画，输入一段文本描述，生成符合描述的人体动画序列。

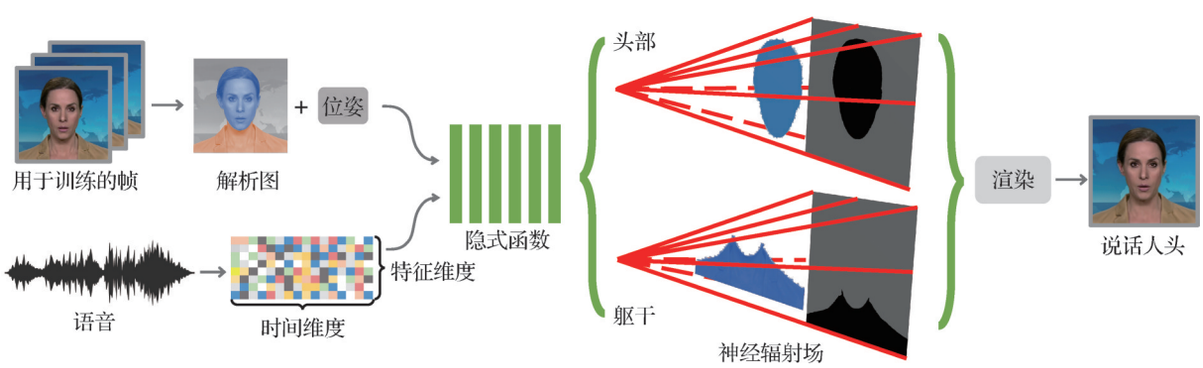


图 6　AD-NeRF 图

Fig. 6　AD-NeRF pipeline diagram

2.1　驱动信号的学习与编码

语音生成人体动画这一任务中，对语音的语义和旋律的学习比较重要。由于语音信号本身采样率很高，通常是先将语音转化成 MFCC 或 Mel 谱，然后作为模型输入。为了进一步建模语音中的语义信息，Rhythmic Gesticulator（Ao 等，2022）基于语言学建模了语音和人体手势的层级特征。低分辨率的特征可以更好地表征语义层面的信息；而高分辨率的

表 1　人头驱动相关方法的比较

Table 1　The comparison of talking head methods

方法	输入	类别	PSNR/dB	优点	缺点
IMAvatar (Zheng等, 2022)	表情系数	Occupancy	23.91～28.75	相比于传统3DMM方法能更好地捕捉几何和外观细节	计算成本高
NeRFBlendShape (Gao等, 2022)	表情系数	NeRF	31.57～36.73	隐式表示的人头模型相比于显式模型渲染更加真实	缺失显式信息，无法直接编辑
NerFACE (Gafni等, 2021)	表情系数	NeRF	23.58～34.94		
HeadNeRF (Hong等, 2022b)	表情系数	NeRF	19.6～21.6		
Wav2Lip (Prajwal等, 2020)	语音	2D	20.17～30.94	速度快、计算资源需求低、易于实现	缺乏三维先验知识和卷积神经网络结构特性，在细节和真实感方面表现不尽人意
MakeItTalk (Zhou等, 2020)	语音	2D	19.2～25.52		
ATVG (Chen等, 2019)	语音	2D	24.13～30.91		
You said that? (Jamaludin等, 2019)	语音	2D	29.91		
Audio2Head (Wang等, 2021b)	语音	2D	21.19～30.93		
DiffTalk (Shen等, 2023)	语音	2D	32.73～34.17		
ADEVP (Ji等, 2021)	语音	2D	29.53		
DialogueNeRF (Zhou等, 2022)	语音	NeRF	24.18～36.15	相比于2D和3D的方法，渲染真实感更高，能够提供从不同视角观察场景的一致性和连续性	需要大量的计算资源和训练数据，训练和推理速度较慢
RAD-NeRF (Tang等, 2022)	语音	NeRF	34		
AD-NeRF (Guo等, 2021b)	语音	NeRF	25.53～31.32		
DFRF (Shen等, 2022)	语音	NeRF	30.75～30.96		
SSP-NeRF (Liu等, 2022)	语音	NeRF	32.65		

注：ATVG为hierarchical cross-modal talking face generation with dynamic pixel-wise loss；ADEVP为audio-driven emotional video portraits；RAD-NeRF为real-time neural talking portrait synthesis。

特征可以表示精细的变化。GestureDiffuCLIP (Ao等, 2023)通过T5 (Raffel等, 2020)提取台本(transcript)词义层面的特征，以此来增强生成结果的语义协同性，并且设计了Video CLIP来支持以视频为参考词(prompt)的风格编辑。

音乐生成舞蹈这一任务中，对音乐的节拍、旋律的提取比较重要。早期工作(Fukayama和Goto, 2015)常常将和弦(chords)、结构(structure)、声谱(spectrogram)、节拍(beats)等作为条件进行推理。ChoreoMaster (Kang等, 2021)将音乐和舞蹈编码到一个统一的编舞隐空间里，以此来提取音乐和舞蹈信号里共性的信息。EDGE(editable dance generation from music)(Tseng等, 2023)没有采用传统的音乐特征，而是使用预训练的音乐特征提取器Jukebox (Dhariwal等, 2020)，取得了更好的效果。

文本生成人体动画这一任务中，侧重于提取文本中与动作相关的语义信息。早期由于数据的缺失和模型能力的限制，一般只是基于运动标签生成人体动画。近年来随着CLIP (Radford等, 2021)的提出，由于其可以提取出文本中与视觉信息相关的部分，所以越来越多的工作开始使用CLIP作为生成模型的输入。也有工作(Jiang等, 2023)直接使用大语言模型提取文本中的语义信息。MotionCLIP (Tevet等, 2022)训练运动自编码器，并保证隐变量与对应的文本/图像的CLIP编码一致，以此来将运动序列编码到CLIP空间下，找出运动和文本语义的对应关系。

2.2　基于RNN的人体动画生成

早期的人体动画生成模型采用RNN结构，Plappert等人(2018)使用Seq2Seq模型(Sutskever等, 2014)学习语言和人体动画之间的对应关系。Text2Action (Ahn等, 2018)在Seq2Seq的基础上，参考了生成对抗模型的设计思路，引入判别器辅助训练。Language2Pose (Ahuja和Morency, 2019)使用RNN模型将文本和运动编码到共同的特征空间之中，然后基于输入的文本，采样出合理的人体运动序列。Ghosh等人(2021)使用门控循环单元(gated recurrent unit, GRU)分别对人体的上半身和下半身

进行了编码，然后进行联合解码得到更真实的人体动画结果。TM2T（Guo 等，2022b）将文本生成人体运动视为翻译任务，用GRU实现了一个神经机器翻译器（neural machine translator）来回归生成人体运动token。Guo 等人（2022a）使用了 GRU+VAE 的模式，进一步提高了模型的生成质量。ChoreoNet（Ye 等，2020）通过用GRU预测编舞运动单元（choreographic action unit）实现舞蹈生成。

RNN模型不善于建模长程信息，且计算效率低、并行性不足，同时扩展性也不理想。目前在人体动画生成任务上有被Transformer模型替代的趋势。

2.3 基于Transformer的人体动画生成

Transformer（Vaswani 等，2017）是一种新型的序列信号学习模型，近年来逐渐取代了RNN模型，其性能优势主要来源于attention机制，具体为

$$Attention\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=f_{\text{softmax}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{8}$$

式中，$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$由词向量通过可学习的线性映射得到，$\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\right)$本质上是对$\boldsymbol{V}$的加权，这样的设计可以避免RNN在处理长序列时的梯度消失或者梯度爆炸问题，同时使模型更好地关注到输入信息中重要的部分。

ACTOR（Petrovich 等，2021）是比较经典的使用Transformer解决人体动画生成问题的工作。如图7所示，其采用了Transformer VAE的结构，将人体序列和动作标签a用Transformer编码为均值μ和方差Σ，之后在训练阶段基于重参数化技巧得到隐变量z，然后解码拟合动作序列，在推理阶段从高斯分布里采样得到z，然后推理动作序列。TEMOS（Petrovich 等，2022）在ACTOR的基础上加入了文本学习分支，使用DistillBERT（Sanh 等，2019）提取文本特征，实现了基于文本的人体运动合成。TEACH（temporal action composition for 3D humans）（Athanasiou 等，2022）在TEMOS的基础上，使用自回归的策略解决了长动画序列在驱动标签切换时不稳定的问题。AI Choreographer（Li 等，2021b）使用了full attention的设计，自回归生成运动序列片段，加强了编舞过程中长舞蹈序列的稳定性。

随着基于Transformer的语言模型如GPT（generative pre-trained Transformer）等在NLP（natural language processing）任务上展现出强大的序列数据处理能力，这些模型也应用到运动序列的生成任务中。这一范式在数学上可以写为

$$p\left(\boldsymbol{S}|c\right)=\Pi_{i=1}^{|S|}p\left(\boldsymbol{S}_i|c,\boldsymbol{S}_{<i}\right) \tag{9}$$

式中，S表示人体运动序列，i是其帧序号。c是某种生成条件，可以是CLIP特征（Zhang 等，2023a），也可以是音乐特征（Li 等，2022）。GPT模型基于$\left(c,\boldsymbol{S}_{<i}\right)$这一信息，自回归预测似然函数$p\left(\boldsymbol{S}_i|c,\boldsymbol{S}_{<i}\right)$。

这一范式的优化目标可以写为

$$L=E_{s\sim p(S)}\left[-\log p\left(\boldsymbol{S}|c\right)\right] \tag{10}$$

T2M-GPT（Zhang 等，2023a）先用 VQ-VAE（van den Oord 等，2017）构建人体运动codebook，然后以CLIP编码的文本特征为条件，自回归生成人体code序列。Bailando（Li 等，2022）使用了强化学习的actor-critic学习的策略来增强舞蹈动画与节拍的一致性。MotionGPT（Jiang 等，2023）将人体运动序列视做外语文本，将人体序列以QA的形式加入到T5模型（Raffel 等，2020）的微调过程中。由于GPT强大的零样本推理能力，MotionGPT可以只用一个模型实现多种人体运动相关任务。

2.4 基于降噪扩散模型的人体动画生成

降噪扩散模型（Ho 等，2020）具有极强的分布建模能力，在图像视频生成、声音合成等领域已经取得了瞩目的成就。现有方法一般是在人体姿态域上构建扩散过程$q\left(x_n|x_{n-1}\right)$，具体为

$$q\left(x_n|x_{n-1}\right)=N\left(x_n;\alpha_n x_{n-1},\beta_n I\right) \tag{11}$$

式中，$n\in 1,\cdots,N$，x_n表示增加噪声程度为n的人体位姿序列信息。α_n表示信号强度，β_n表示加噪程度。这一扩散过程将数据分布逐渐映射为高斯白噪声，进而有

$$p\left(x_N\right)=N\left(x_N;0,I\right) \tag{12}$$

扩散过程的逆过程用高斯分布近似写为

$$p\left(x_{n-1}|x_n\right)=N\left(\mu\left(x_n,n\right),\Sigma\left(x_n,n\right)\right) \tag{13}$$

一般地，$\Sigma\left(x_n,n\right)$直接用单位阵来表示，降噪扩散模型就是用神经网络来预测$\mu\left(x_n,n\right)$，Ho 等人（2020）指出这可以归结于学习降噪网络ϵ_θ。

降噪扩散模型的优化目标为

$$L=E_{x_0,n,\epsilon}\left[k_n\|\epsilon-\epsilon_\theta\left(x_n,n\right)\|\right] \tag{14}$$

式中，k_n是对不同噪声程度下降噪loss的权重，ϵ_θ是降噪网络。

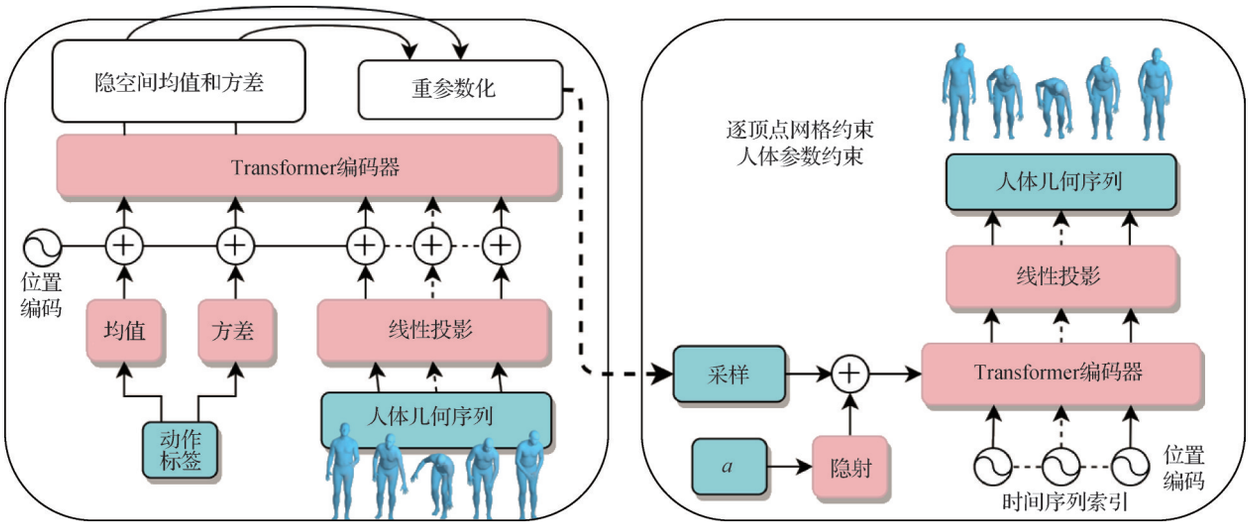


图7　基于Transformer VAE的ACTOR管线图

Fig. 7　ACTOR pipeline diagram based on Transformer VAE

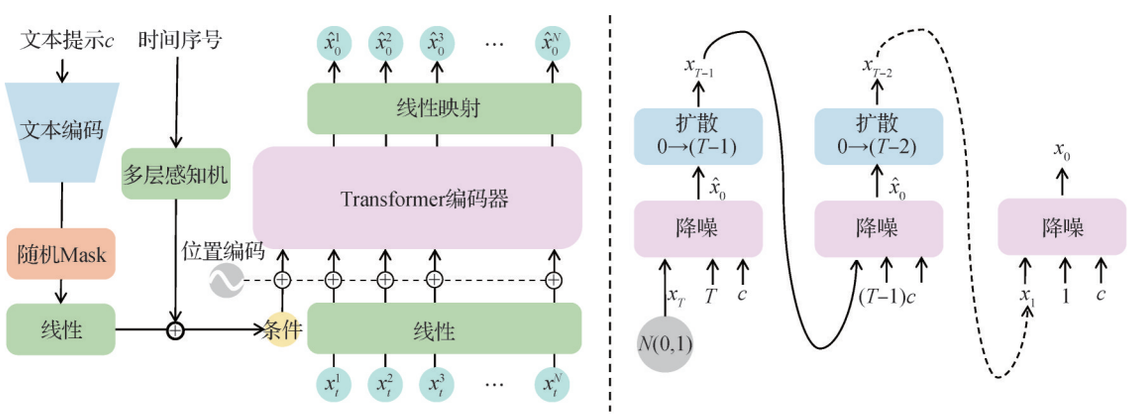
FLAME(Kim等,2023)、MDM(Tevet等,2023)和MotionDiffuse(Zhang等,2022)是最早采用降噪扩散模型实现人体动画生成的工作,都取得了比较理想的效果。图8展示的是MDM方法,其通过Transformer实现基于CLIP特征和t的序列降噪,通过降噪过程将高斯分布变为运动序列分布。MoFusion(Dabral等,2023)使用一种基于多模态Transformer的1D U-Net实现降噪网络,可以从音频或文本输入中生成逼真的3D人体运动序列,具有高度的可塑性和真实感。

GestureDiffuCLIP(Ao等,2023)使用多模态的CLIP编码以及AdaIN(Huang和Belongie,2017)的设计思路,实现了多模态的人体动画风格编辑。该方法支持多模态的提示(如文本、运动序列和视频)来控制手势风格,但是生成的动作中存在足部滑动问题。

EDGE(Tseng等,2023)根据音乐条件生成多样化、物理上合理的舞蹈编排,并且具有强大的编辑能力。EDGE方法通过使用Jukebox进行音乐特征提取,能够处理多种类型的音乐,设计了Contact Consistency Loss来解决动画生成过程中的滑步问题。但是在高强度的舞蹈动作中,滑步问题仍然存在。

2.5　其他方法

也有一些工作采用了其他思路来处理人体运动生成的问题,Learn2Dance(Ofli等,2011)使用隐式马尔可夫模型实现人体动画的生成。Fan等人(2012)先构建人体运动motion graph,然后通过动态规划求解最优舞蹈路径。TalkSHOW(Yi等,2023)参考



(a) MDM的降噪网络　　(b) MDM的采样过程

图8　MDM的降噪网络和采样过程

Fig. 8　Denoising network and sampling process of MDM ((a) denoising network of MDM; (b) sampling process of MDM)

MeshTalk(Richard 等,2021),将自回归地在类别隐空间生成的思路借鉴过来,构建人体动画类别隐空间然后回归。ChoreoMaster(Kang 等,2021)先构建了编舞编码统一隐空间,然后在这样的空间里进行图优化(graph optimization),也取得了很好的效果。

3 多模态数字人形象构建

3.1 视觉语言相似性引导的虚拟形象构建

早期工作使用生成对抗网络实现人像编辑(曹申豪 等,2022)(廖远鸿 等,2023)(王艺博 等,2023),CLIP(Radford 等,2021)的提出搭建了视觉和语言的桥梁,也使得基于文本编辑和生成虚拟形象成为可能。在 CLIP 的隐空间中,定义前后文本编码差异 ΔT 和图像编码差异 ΔI,具体为

$$\begin{cases}\Delta T = E_T(\boldsymbol{t}_{\text{target}}) - E_T(\boldsymbol{t}_{\text{source}}) \\ \Delta I = E_I(\boldsymbol{I}_{\text{target}}) - E_I(\boldsymbol{I}_{\text{source}})\end{cases} \tag{15}$$

式中,$\boldsymbol{t}_{\text{target}}$ 和 $\boldsymbol{t}_{\text{source}}$ 为目标文本和源文本,$\boldsymbol{I}_{\text{target}}$ 和 $\boldsymbol{I}_{\text{source}}$ 为目标图像和源图像,E_T 和 E_I 分别为 CLIP 的图像和文本编码器。CLIP 引导 loss 函数可以通过差异向量的余弦相似性定义为

$$L_{\text{dir}} = 1 - \frac{\Delta T \cdot \Delta I}{|\Delta T||\Delta I|} \tag{16}$$

CLIPstyler(Kwon 和 Ye, 2022)采用逐 patch 的 CLIP 损失函数来引导图像风格化网络的学习,使得风格化后的图像符合文本描述。StyleCLIP(Patashnik 等,2021)在预训练好的 StyleGAN(Karras 等,2020)的基础上,提出了隐变量优化法(latent optimization)、隐空间映射器(latent mapper)学习法、全局方向(global directions)搜索法,平衡了生成质量和构建效率的矛盾。StyleGAN-NADA(Gal 等,2021)给出了基于文本描述实现 StyleGAN 域迁移效果的方法,先用预训练 StyleGAN 做 $W+$ 空间的逆转,然后选择隐变量变化显著的层进行微调,这种只微调原本网络一部分的思路有效克服了模式坍塌(mode collapse)和过拟合(overfitting),取得了比较好的效果。图 9 展示了 StyleGAN-NADA 如何使用 CLIP 空间下的文本方向向量 ΔT 引导图像生成器优化。参考 CLIP 中定义,引入了两个图像生成器 G_{frozen} 和 G_{train},G_{frozen} 是在整个训练过程中保持冻结的生成器,其主要作用是为每个潜在代码提供源域中的图像,G_{train} 是经过微调的生成器,其训练目标是在文本描述的方向上产生与源生成器略有不同的图像。然后在式(15)中将 $\boldsymbol{I}_{\text{target}}$ 替换为 $\boldsymbol{G}_{\text{train}}(w)$,将 $\boldsymbol{I}_{\text{source}}$ 替换为 $G_{\text{frozen}}(w)$。

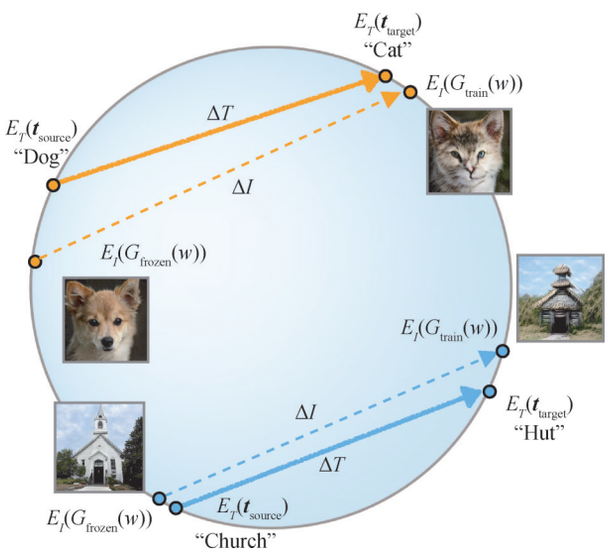


图 9 CLIP 空间下的文本方向向量引导图像生成器优化

Fig. 9 The text direction vector in CLIP space guides the optimization of the image generator

也有工作将多模态信号映射到人头生成模型的隐空间下,以此来实现多模态的人像生成。TediGAN 用 StyleGAN(Karras 等,2020)将真实人头图像逆转(inversion)到隐空间下,然后基于视觉语言相似性将文本也映射到同一空间中。基于图像隐变量 w^v 和语言隐变量 w^l 进行风格混合,最后解码得到生成结果。

3.2 多模态降噪扩散模型引导的虚拟形象构建

随着 Imagen(Saharia 等,2022)、DALL-E2 (Ramesh 等,2022)和 stable diffusion (Rombach 等,2021)等降噪扩散模型 (Ho 等,2020)在二维图像生成领域获得巨大成功,也有工作考虑用这样的二维图像生成模型引导数字形象的构建。

大部分这样的工作都基于 DreamFusion (Poole 等,2022)提出的 SDS(score distillation sampling)来处理这一问题。如图 10 所示,设图像的参数表示为 $x = g(\theta)$,则对其进行扩散后,降噪网络输出分数与真实噪声的差异为

$$L_{\text{diff}} = E_{t\sim U(0,1),\epsilon\sim N(0,I)}[\, w(t) \| \epsilon_\phi(\alpha_t x + \sigma_t \epsilon\,;\, t) - \epsilon \|_2^2] \tag{17}$$

式中，$w(t)$是依赖于t的权重函数，其余可参考 Poole 等人（2022）的说明，L_{diff}的梯度为

$$\nabla_\theta L_{\text{diff}} = E_{t,\epsilon}\left[w(t)\left(\epsilon_\phi\left(\hat{z}_t; y, t\right) - \epsilon\right) \cdot \frac{\partial \hat{\epsilon}_\phi\left(z_t; y, t\right)}{z_t} \frac{\partial x}{\partial \theta} \right] \tag{18}$$

由于$\frac{\partial \hat{\epsilon}_\phi\left(z_t; y, t\right)}{z_t}$计算量巨大，忽略后计算近似梯度，进而得到

$$\nabla_\theta L_{\text{SDS}} = E_{t,\epsilon}\left[w(t)\left(\hat{\epsilon}_\phi\left(z_t; y, t\right) - \epsilon\right) \frac{\partial x}{\partial \theta} \right] \tag{19}$$

利用这个梯度，对随机初始化的3D模型进行优化，使其从任意角度呈现低误差的2D渲染，进而生成3D模型，实现了高质量的3D合成。DreamFusion中使用预训练的图像扩散模型，从而不需要大规模的3D数据集对模型进行修改。其参数化图像是一种NeRF表示。

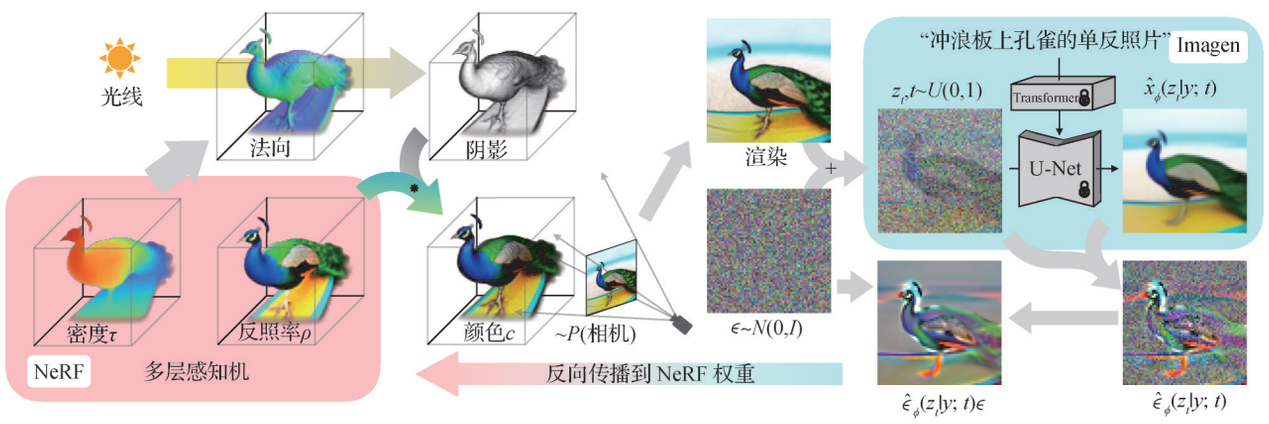


图10　SDS借助多模态降噪扩散模型（如Imagen）引导NeRF训练

Fig. 10　SDS utilizes a multi-modal denoising diffusion model, such as Imagen, to guide NeRF training

Instruct-NeRF2NeRF（Haque 等，2023）使用 Instruct-Pix2Pix（Brooks 等，2023）来提取形状和外观先验，通过纯文本指令来编辑3D NeRF场景。该方法的编辑方式直观高效，编辑范围广，并且通过在预捕获的NeRF场景上进行编辑，确保了生成的场景具有三维一致性。不足之处在于依赖InstructPix2Pix的表达能力，需要预捕捉的NeRF场景，优化过程较长。

DreamAvatar（Cao 等，2023）和 DreamWaltz（Huang 等，2023）通过结合人体 SMPL（skinned multi-person linear）先验，实现了高质量的3D角色创建。此外，该方法还学习了一个可动画化的NeRF表示，可以将生成的角色转移到任意姿势，实现逼真的动画效果。

AvatarCLIP（Hong 等，2022a）和 AvatarBooth（Zeng 等，2023）结合了 NeuS（Wang 等，2021a）的几何表示，实现了多模态高质量人体建模。AvatarCLIP 结合了 CLIP 和 SMPL 等技术，能够通过自然语言描述生成未见过的3D角色，并能够自动化地为其添加动画效果。AvatarBooth 根据文本提示或图像集生成高质量的、可定制的3D人物形象。

DreamFace（Zhang 等，2023b）在 ICT-FaceKit（Li 等，2020）的基础上，借助SDS使用文本优化了几何和纹理，实现一种文本驱动个性化3D人脸生成框架。该框架能够根据用户提供的文本描述生成高度个性化、逼真的3D人脸模型，并支持丰富的动画表现。具体而言，其采用了一种逐步生成的方法，将文本描述转化为3D人脸模型的几何形状、纹理和动画能力等特征，并通过优化算法得到最终的生成结果。该框架不仅具有高度的可定制性和可扩展性，而且能够轻松地嵌入到现有的计算机图形生成流程中。

3.3　三维多模态虚拟人生成模型

也有工作致力于直接构建三维多模态生成模型。这一类问题的难点在于三维数据集的获取。DATID-3D（Kim 和 Chun，2023）先采用 shallow diffusion 的策略将 EG3D（efficient geometry-aware 3D GANs）（Chan 等，2022）生成的人头图像根据提示词（prompt）进行风格化，然后筛选出符合需求的图像样本构建数据集，最后将这样的数据用于EG3D的训练，得到三维风格化生成模型。Rodin（Wang 等，2023）则在三维空间上用 triplane（Chan 等，2022）构造以文本为 prompt 的生成模型，使用图形引擎生成的三维虚拟人数据进行训练。

4 结 语

4.1 问题分析与总结

本文分别对多模态人头动画、多模态人体动画和多模态数字人形象构建的相关技术进行了分类，并介绍了对应的代表性工作，现对本文讨论内容加以总结：

1）人头模型表达能力和渲染质量。在表情驱动人头的任务中，显式模型使用简单、可编辑性强、计算高效，但是模型表达能力弱，渲染不具有可微性，且在建模人脸个性化特征以及表示非人脸区域（如头发等）上困难比较多；隐式模型尤其是NeRF表示建模能力更强，渲染更加真实。

2）语音驱动人头的生成质量。在语音驱动人头的任务中，二维说话人方案由于三维先验知识的缺乏以及CNN的固有缺陷，生成质量往往不够理想；语音驱动三维数字人头几何动画的方案依赖监督数据质量，且动画效果受限于参数化网格表达能力；语音驱动NeRF人头方案可以克服上述问题，可以生成更自然真实的说话人视频。

3）人体动画模型的表达能力。在多模态人体动画的任务中，RNN方法由于不善于捕捉长程信息、计算效率低等缺陷，已经渐渐被Transformer模型替代，同时降噪扩散模型由于其具有极强的分布建模能力，非常适合解决多模态人体动画中一到多的映射难题，也在相关问题上展示出很强的潜力。

4）数字人形象构建的挑战。在多模态数字人形象构建任务中，视觉语言相似性引导的虚拟形象构建方法依靠CLIP模型的相似性损失函数，引导了模型或参数的优化。而基于多模态降噪扩散模型引导的方法依靠SDS，利用现有多模态图像生成模型的知识，引导三维形象的构建。这两种方案都能得到较好的效果。三维多模态虚拟人的直接构建面临挑战，尽管有视觉语言相似性和降噪扩散模型引导的方法存在，但问题仍然是缺乏足够的三维虚拟人数据集。

4.2 未来展望

以下给出了多模态数字人领域未来的重点研究方向，其中技术进步与社会影响同样重要。未来研究将不仅致力于提高技术水平和表现能力，同时也会更加注重伦理、法律和社会责任等方面的考量。

1）高精度三维建模与实时渲染。为了能准确地捕捉人体表面和特征的细微变化，改进三维建模技术也是至关重要的。这可能包括对复杂区域（如头发、眼睛等）更精细的渲染和建模，以提高数字人的真实感和个性化程度。当前基于神经体积渲染（NeRF）等隐式表示方法是一个重要的研究方向，以便更好地捕捉光线与表面相互作用的细微差异。而近期提出的3D-Gaussian（Kerbl等，2023）模型在显式建模三维场景的前提下同时具有高精度与高渲染速率的优势，使用3D-Gaussian进行数字人建模也是当前的研究热点。

2）大规模数据集与生成模型。构建更大规模、更多样化的数据集将是未来一个重要挑战。利用这些数据集，研究人员可以训练更强大、更通用的生成模型，能够适应各种不同的语境和表现形式，从而提高数字人模型的泛化能力和适用性。

3）多模态融合与跨模态学习。将探索多种信息源的融合，例如语音、图像和文本等，也是一个重要的研究方向。跨模态学习可以让模型从不同的数据来源中学习，并在各种任务中展现更强大的性能。这可能导致更丰富、更全面的数字人模型，能够在不同场景和需求下展现出更多样化的行为和表现。

4）伦理和社会影响的研究。随着数字人技术的发展，未来研究也将更加关注其在社会和伦理层面的影响。这包括对数字身份、欺骗性用途和隐私问题的关注，以及数字人在辅助功能、教育和娱乐等领域带来的潜在影响。

以上研究方向可能引领多模态数字人领域朝着更全面、逼真和通用的方向发展。数字人作为元宇宙的重要组成部分，其发展将在很大程度上影响未来虚拟空间的交互和体验。

参考文献（Reference）

Ahn H, Ha T, Choi Y, Yoo H and Oh S. 2018. Text2Action: generative adversarial synthesis from language to action//Proceedings of 2018 IEEE International Conference on Robotics and Automation (ICRA). Brisbane, Australia: IEEE: 5915-5920 [DOI: 10.1109/ICRA.2018.8460608]

Ahuja C and Morency L P. 2019. Language2Pose: natural language grounded pose forecasting//Proceedings of 2019 International Conference on 3D Vision (3DV). Quebec City, Canada: IEEE: 719-728 [DOI: 10.1109/3DV.2019.00084]

Ao T L, Gao Q Z, Lou Y K, Chen B Q and Liu L B. 2022. Rhythmic gesticulator: rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. ACM Transactions on Graphics, 41(6): #209 [DOI: 10.1145/3550454.3555435]

Ao T L, Zhang Z Y and Liu L B. 2023. GestureDiffuCLIP: gesture diffusion model with CLIP latents. ACM Transactions on Graphics, 42(4) [DOI: 10.1145/3592097]

Athanasiou N, Petrovich M, Black M J and Varol G. 2022. TEACH: temporal action composition for 3D humans//Proceedings of 2022 International Conference on 3D Vision (3DV). Prague, Czech Republic: IEEE: 414-423 [DOI: 10.1109/3DV57658.2022.00053]

Blanz V and Vetter T. 1999. A morphable model for the synthesis of 3D faces//Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. Los Angeles, USA: ACM: 187-194 [DOI: 10.1145/311535.311556]

Brooks T, Holynski A and Efros A A. 2023. InstructPix2Pix: learning to follow image editing instructions//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 18392-18402 [DOI: 10.1109/CVPR52729.2023.01764]

Cao C, Weng Y L, Zhou S, Tong Y Y and Zhou K. 2014. FaceWarehouse: a 3D facial expression database for visual computing. IEEE Transactions on Visualization and Computer Graphics, 20(3): 413-425 [DOI: 10.1109/TVCG.2013.249]

Cao S H, Liu X H, Mao X Q and Zou Q. 2022. A review of human face forgery and forgery-detection technologies. Journal of Image and Graphics, 27(4): 1023-1038 (曹申豪, 刘晓辉, 毛秀青, 邹勤. 2022. 人脸伪造及检测技术综述. 中国图象图形学报, 27(4): 1023-1038) [DOI: 10.11834/jig.200466]

Cao Y K, Cao Y P, Han K, Shan Y and Wong K Y K. 2023. DreamAvatar: text-and-shape guided 3D human avatar generation via diffusion models [EB/OL]. [2023-08-20]. https://arxiv.org/pdf/2304.00916.pdf

Chan E R, Lin C Z, Chan M A, Nagano K, Pan B X, de Mello S, Gallo O, Guibas L, Tremblay J, Khamis S, Karras T and Wetzstein G. 2022. Efficient geometry-aware 3D generative adversarial networks//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 16102-16112 [DOI: 10.1109/CVPR52688.2022.01565]

Chen L L, Maddox R K, Duan Z Y and Xu C L. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 7824-7833 [DOI: 10.1109/cvpr.2019.00802]

Chu B, Romdhani S and Chen L M. 2014. 3D-aided face recognition robust to expression and pose variations//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 1907-1914 [DOI: 10.1109/CVPR.2014.245]

Chung J S and Zisserman A. 2016. Out of time: automated lip sync in the wild//Proceedings of ACCV 2016 International Workshops on Computer Vision. Taipei, China: Springer: 251-263 [DOI: 10.1007/978-3-319-54427-4_19]

Cudeiro D, Bolkart T, Laidlaw C, Ranjan A and Black M J. 2019. Capture, learning, and synthesis of 3D speaking styles//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 10093-10103 [DOI: 10.1109/CVPR.2019.01034]

Dabral R, Mughal M H, Golyanik V and Theobalt C. 2023. MoFusion: a framework for denoising-diffusion-based motion synthesis//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 9760-9770 [DOI: 10.1109/CVPR52729.2023.00941]

Dhariwal P, Jun H, Payne C, Kim J W, Radford A and Sutskever I. 2020. Jukebox: a generative model for music [EB/OL]. [2023-08-20]. https://arxiv.org/pdf/2005.00341.pdf

Edwards P, Landreth C, Fiume E and Singh K. 2016. JALI: an animator-centric viseme model for expressive lip synchronization. ACM Transactions on Graphics, 35(4): #127 [DOI: 10.1145/2897824.2925984]

Fan R K, Xu S H and Geng W D. 2012. Example-based automatic music-driven conventional dance motion synthesis. IEEE Transactions on Visualization and Computer Graphics, 18(3): 501-515 [DOI: 10.1109/TVCG.2011.73]

Fan Y R, Lin Z J, Saito J, Wang W P and Komura T. 2021. FaceFormer: speech-driven 3D facial animation with Transformers//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 18749-18758 [DOI: 10.1109/CVPR52688.2022.01821]

Fisher C G. 1968. Confusions among visually perceived consonants. Journal of Speech and Hearing Research, 11(4): 796-804 [DOI: 10.1044/jshr.1104.796]

Fukayama S and Goto M. 2015. Music content driven automated choreography with beat-wise motion connectivity constraints//Proceedings of the 12th Sound and Music Computing Conference. Maynooth, Ireland: [s.n.]: 177-183

Gafni G, Thies J, Zollhöfer M and Nießner M. 2021. Dynamic neural radiance fields for monocular 4D facial avatar reconstruction//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 8645-8654 [DOI: 10.1109/CVPR46437.2021.00854]

Gal R, Patashnik O, Maron H, Chechik G and Cohen-Or D. 2021. STYLEGAN-NADA: CLIP-guided domain adaptation of image generators [EB/OL]. [2023-08-20]. https://arxiv.org/pdf/2108.00946.pdf

Gao X, Zhong C L, Xiang J, Hong Y, Guo Y D and Zhang J Y. 2022. Reconstructing personalized semantic facial NeRF models from monocular video. ACM Transactions on Graphics, 41(6): #200

[DOI: 10.1145/3550454.3555501]

Ghosh A, Cheema N, Oguz C, Theobalt C and Slusallek P. 2021. Synthesis of compositional animations from textual descriptions//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 1376-1386 [DOI: 10.1109/ICCV48922.2021.00143]

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial nets//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 2672-2680

Guo C, Zou S H, Zuo X X, Wang S, Ji W, Li X Y and Cheng L. 2022a. Generating diverse and natural 3D human motions from text//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 5142-5151 [DOI: 10.1109/CVPR52688.2022.00509]

Guo C, Zuo X X, Wang S and Cheng L. 2022b. TM2T: stochastic and tokenized modeling for the reciprocal generation of 3D human motions and texts//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 580-597 [DOI: 10.1007/978-3-031-19833-5_34]

Guo Y D, Cai L and Zhang J Y. 2021a. 3D face from X: learning face shape from diverse sources. IEEE Transactions on Image Processing, 30: 3815-3827 [DOI: 10.1109/TIP.2021.3065798]

Guo Y D, Chen K Y, Liang S, Liu Y J, Bao H J and Zhang J Y. 2021b. AD-NeRF: audio driven neural radiance fields for talking head synthesis//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 5764-5774 [DOI: 10.1109/ICCV48922.2021.00573]

Hannun A, Case C, Casper J, Catanzaro B, Diamos G, Elsen E, Prenger R, Satheesh S, Sengupta S, Coates A and Ng A Y. 2014. Deep speech: scaling up end-to-end speech recognition [EB/OL]. [2023-08-20]. https://arxiv.org/pdf/1412.5567.pdf

Haque A, Tancik M, Efros A A, Holynski A and Kanazawa A. 2023. Instruct-NeRF2NeRF: editing 3D scenes with instructions//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 19683-19693 [DOI: 10.1109/ICCV51070.2023.01808]

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #574

Hong F Z, Zhang M Y, Pan L, Cai Z A, Yang L and Liu Z W. 2022a. AvatarCLIP: zero-shot text-driven generation and animation of 3D avatars. ACM Transactions on Graphics, 41(4): #161 [DOI: 10.1145/3528223.3530094]

Hong Y, Peng B, Xiao H Y, Liu L G and Zhang J Y. 2022b. HeadNeRF: a realtime NeRF-based parametric head model//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 20342-20352 [DOI: 10.1109/CVPR52688.2022.01973]

Huang X and Belongie S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 1510-1519 [DOI: 10.1109/iccv.2017.167]

Huang Y K, Wang J N, Zeng A L, Cao H, Qi X B, Shi Y K, Zha Z J and Zhang L. 2023. DreamWaltz: make a scene with complex 3D animatable avatars//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: NeurIPS

Isola P, Zhu J Y, Zhou T H and Efros A A. 2017. Image-to-image translation with conditional adversarial networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 5967-5976 [DOI: 10.1109/CVPR.2017.632]

Jamaludin A, Chung J S and Zisserman A. 2019. You said that?: synthesising talking faces from audio. International Journal of Computer Vision, 127 (11): 1767-1779 [DOI: 10.1007/s11263-019-01150-y]

Ji X Y, Zhou H, Wang K S Y, Wu W, Loy C C, Cao X and Xu F. 2021. Audio-driven emotional video portraits//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: 14075-14084 [DOI: 10.1109/CVPR46437.2021.01386]

Jiang B, Chen X, Liu W, Yu J Y, Yu G and Chen T. 2023. MotionGPT: human motion as a foreign language//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: NeurIPS

Kang C, Tan Z P, Lei J, Zhang S H, Guo Y C, Zhang W D and Hu S M. 2021. ChoreoMaster: choreography-oriented music-driven dance synthesis. ACM Transactions on Graphics, 40(4): #145 [DOI: 10.1145/3450626.3459932]

Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J and Aila T. 2020. Analyzing and improving the image quality of styleGAN//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 8107-8116 [DOI: 10.1109/CVPR42600.2020.00813]

Kerbl B, Kopanas G, Leimkuehler T and Drettakis G. 2023. 3D Gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics, 42(4): #139 [DOI: 10.1145/3592433]

Kim G and Chun S Y. 2023. DATID-3D: diversity-preserved domain adaptation using text-to-image diffusion for 3D generative model//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 14203-14213 [DOI: 10.1109/CVPR52729.2023.01365]

Kim J, Kim J and Choi S. 2023. FLAME: free-form language-based motion synthesis & editing//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI: 8255-

8263［DOI：10.1609/aaai.v37i7.25996］

Kwon G and Ye J C. 2022. CLIPstyler：image style transfer with a single text condition//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. New Orleans，USA：IEEE：18041-18050［DOI：10.1109/CVPR52688.2022.01753］

Li R L，Bladin K，Zhao Y J，Chinara C，Ingraham O，Xiang P D，Ren X L，Prasad P，Kishore B，Xing J and Li H. 2020. Learning formation of physically-based face attributes//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. Seattle，USA：IEEE：3407-3416［DOI：10.1109/CVPR42600.2020.00347］

Li R L，Yang S，Ross D A and Kanazawa A. 2021b. AI choreographer：music conditioned 3D dance generation with AIST++//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal，Canada：IEEE：13381-13392［DOI：10.1109/ICCV48922.2021.01315］

Li S Y，Yu W J，Gu T P，Lin C Z，Wang Q，Qian C，Loy C C and Liu Z W. 2022. Bailando：3D dance generation by actor-critic GPT with choreographic memory//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. New Orleans，USA：IEEE：11040-11049［DOI：10.1109/CVPR52688.2022.01077］

Li T Y，Bolkart T，Black M J，Li H and Romero J. 2017. Learning a model of facial shape and expression from 4D scans. ACM Transactions on Graphics，36(6)：#194［DOI：10.1145/3130800.3130813］

Liao Y H，Qian W H and Cao J D. 2023. MStarGAN：a face style transfer network with changeable style intensity. Journal of Image and Graphics，28(12)：3784-3796（廖远鸿，钱文华，曹进德. 2023. 风格强度可变的人脸风格迁移网络. 中国图象图形学报，28(12)：3784-3796）［DOI：10.11834/jig.221149］

Liu X，Xu Y H，Wu Q Y，Zhou H，Wu W and Zhou B L. 2022. Semantic-aware implicit neural audio-driven video portrait generation//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv，Israel：Springer：106-125［DOI：10.1007/978-3-031-19836-6_7］

Medsker L R and Jain L. 2001. Recurrent neural networks. Design and Applications，5：64-67

Mildenhall B，Srinivasan P P，Tancik M，Barron J T，Ramamoorthi R and Ng R. 2020. NeRF：representing scenes as neural radiance fields for view synthesis//Proceedings of the 16th European Conference on Computer Vision. Glasgow，UK：Springer：405-421［DOI：10.1007/978-3-030-58452-8_24］

Ofli F，Erzin E，Yemez Y and Tekalp A M. 2011. Learn2Dance：learning statistical music-to-dance mappings for choreography synthesis. IEEE Transactions on Multimedia，14(3)：747-759［DOI：10.1109/TMM.2011.2181492］

Patashnik O，Wu Z Z，Shechtman E，Cohen-Or D and Lischinski D. 2021. StyleCLIP：text-driven manipulation of StyleGAN imagery//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision（ICCV）. Montreal，Canada：IEEE：2065-2074［DOI：10.1109/ICCV48922.2021.00209］

Petrovich M，Black M J and Varol G. 2021. Action-conditioned 3D human motion synthesis with Transformer VAE//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal，Canada：IEEE：10965-10975［DOI：10.1109/ICCV48922.2021.01080］

Petrovich M，Black M J and Varol G. 2022. TEMOS：generating diverse human motions from textual descriptions//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv，Israel：Springer：480-497［DOI：10.1007/978-3-031-20047-2_28］

Plappert M，Mandery C and Asfour T. 2018. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. Robotics and Autonomous Systems，109：13-26［DOI：10.1016/j.robot.2018.07.006］

Poole B，Jain A，Barron J T and Mildenhall B. 2022. DreamFusion：text-to-3D using 2D diffusion//Proceedings of the 11th International Conference on Learning Representations. Kigali，Rwanda：ICLR

Prajwal K R，Mukhopadhyay R，Namboodiri V P and Jawahar C V. 2020. A lip sync expert is all you need for speech to lip generation in the wild//Proceedings of the 28th ACM International Conference on Multimedia. Seattle，USA：ACM：484-492［DOI：10.1145/3394171.3413532］

Radford A，Kim J W，Hallacy C，Ramesh A，Goh G，Agarwal S，Sastry G，Askell A，Mishkin P，Clark J，Krueger G and Sutskever I. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning.［s.l.］：PMLR：8748-8763

Raffel C，Shazeer N，Roberts A，Lee K，Narang S，Matena M，Zhou Y Q，Li W and Liu P J. 2020. Exploring the limits of transfer learning with a unified text-to-text Transformer. The Journal of Machine Learning Research，21(1)：#140

Ramesh A，Dhariwal P，Nichol A，Chu C and Chen M. 2022. Hierarchical text-conditional image generation with CLIP latents［EB/OL］.［2023-08-20］. https://arxiv.org/pdf/2204.06125.pdf

Ranjan A，Bolkart T，Sanyal S and Black M J. 2018. Generating 3D faces using convolutional mesh autoencoders//Proceedings of the 15th European Conference on Computer Vision. Munich，Germany：Springer：725-741［DOI：10.1007/978-3-030-01219-9_43］

Richard A，Zollhöfer M，Wen Y D，de la Torre F and Sheikh Y. 2021. MeshTalk：3D face animation from speech using cross-modality disentanglement//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision（ICCV）. Montreal，Canada：IEEE：1153-1162［DOI：10.1109/ICCV48922.2021.00121］

Rombach R，Blattmann A，Lorenz D，Esser P and Ommer B. 2021. High-resolution image synthesis with latent diffusion models//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans，USA：IEEE：10674-10685

[DOI: 10.1109/CVPR52688.2022.01042]

Saharia C, Chan W, Saxena S, Li L L, Whang J, Denton E L, Ghasemipour K, Gontijo Lopes R, Karagol Ayan B, Salimans T, Ho J, Fleet D J and Norouzi M. 2022. Photorealistic text-to-image diffusion models with deep language understanding//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #2643

Sanh V, Debut L, Chaumond J and Wolf T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter [EB/OL]. [2023-08-20]. https://arxiv.org/pdf/1910.01108.pdf

Shen S, Li W H, Zhu Z, Duan Y Q, Zhou J and Lu J W. 2022. Learning dynamic facial radiance fields for few-shot talking head synthesis//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 666-682 [DOI: 10.1007/978-3-031-19775-8_39]

Shen S, Zhao W L, Meng Z B, Li W H, Zhu Z, Zhou J and Lu J W. 2023. DiffTalk: crafting diffusion models for generalized audio-driven portraits animation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 1982-1991 [DOI: 10.1109/CVPR52729.2023.00197]

Sun J X, Wang X, Wang L Z, Li X Y, Zhang Y, Zhang H W and Liu Y B. 2023. Next3D: generative neural texture rasterization for 3D-aware head avatars//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 20991-21002 [DOI: 10.1109/CVPR52729.2023.02011]

Sutskever I, Vinyals O and Le Q V. 2014. Sequence to sequence learning with neural networks//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 3104-3112

Tang J X, Wang K S Y, Zhou H, Chen X K, He D L, Hu T S, Liu J T, Zeng G and Wang J D. 2022. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition [EB/OL]. [2023-08-20]. https://arxiv.org/pdf/2211.12368.pdf

Tevet G, Gordon B, Hertz A, Bermano A H and Cohen-Or D. 2022. MotionCLIP: exposing human motion generation to clip space//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 358-374 [DOI: 10.1007/978-3-031-20047-2_21]

Tevet G, Raab S, Gordon B, Shafir Y, Cohen-or D and Bermano A H. 2023. Human motion diffusion model//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: ICLR

Thies J, Elgharib M, Tewari A, Theobalt C and Nießner M. 2020. Neural voice puppetry: audio-driven facial reenactment//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 716-731 [DOI: 10.1007/978-3-030-58517-4_42]

Tran L and Liu X M. 2018. Nonlinear 3D face morphable model//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7346-7355 [DOI: 10.1109/CVPR.2018.00767]

Tseng J, Castellon R and Liu C K. 2023. EDGE: editable dance generation from music//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 448-458 [DOI: 10.1109/CVPR52729.2023.00051]

van den Oord A, Kalchbrenner N, Vinyals O, Espeholt L, Graves A and Kavukcuoglu K. 2016. Conditional image generation with PixelCNN decoders//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc.: 4797-4805

van den Oord A, Vinyals O and Kavukcuoglu K. 2017. Neural discrete representation learning//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6309-6318

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Wang P, Liu L J, Liu Y, Theobalt C, Komura T and Wang W P. 2021a. NeuS: learning neural implicit surfaces by volume rendering for multi-view reconstruction//Proceedings of the 35th International Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS: 27171-27183

Wang S Z, Li L C, Ding Y, Fan C J and Yu X. 2021b. Audio2Head: audio-driven one-shot talking-head generation with natural head motion//Proceedings of the 30th International Joint Conference on Artificial Intelligence. Montreal, Canada: IJCAI: 1098-1105 [DOI: 10.24963/ijcai.2021/152]

Wang T F, Zhang B, Zhang T, Gu S Y, Bao J M, Baltrusaitis T, Shen J J, Chen D, Wen F, Chen Q F and Guo B M. 2023. RODIN: a generative model for sculpting 3D digital avatars using diffusion//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 4563-4573 [DOI: 10.1109/CVPR52729.2023.00443]

Wang Y B, Zhang K, Kong Y H, Yu T T and Zhao S W. 2023. Overview of human-facial-related age syntheis based generative adversarial network methods. Journal of Image and Graphics, 28(10): 3004-3024 (王艺博, 张珂, 孔英会, 于婷婷, 赵士玮. 2023. 人脸年龄合成的生成对抗网络方法综述. 中国图象图形学报, 28(10): 3004-3024) [DOI: 10.11834/jig.220842]

Xing J B, Xia M H, Zhang Y C, Cun X D, Wang J and Wong T T. 2023. CodeTalker: speech-driven 3D facial animation with discrete motion prior//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 12780-12790 [DOI: 10.1109/CVPR52729.2023.01229]

Ye Z H, Jiang Z Y, Ren Y, Liu J L, He J Z and Zhao Z. 2023. Gene-

Face: generalized and high-fidelity audio-driven 3D talking face synthesis//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: ICLR

Ye Z J, Wu H Z, Jia J, Bu Y H, Chen W, Meng F B and Wang Y F. 2020. ChoreoNet: towards music to dance synthesis with choreographic action unit//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 744-752 [DOI: 10.1145/3394171.3414005]

Yenamandra T, Tewari A, Bernard F, Seidel H P, Elgharib M, Cremers D and Theobalt C. 2021. i3DMM: deep implicit 3D morphable model of human heads//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 12798-12808 [DOI: 10.1109/CVPR46437.2021.01261]

Yi H W, Liang H L, Liu Y F, Cao Q, Wen Y D, Bolkart T, Tao D C and Black M J. 2023. Generating holistic 3D human motion from speech//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 469-480 [DOI: 10.1109/CVPR52729.2023.00053]

Zakharov E, Shysheya A, Burkov E and Lempitsky V. 2019. Few-shot adversarial learning of realistic neural talking head models//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 9458-9467 [DOI: 10.1109/ICCV.2019.00955]

Zeng Y F, Lu Y X, Ji X Y, Yao Y, Zhu H and Cao X. 2023. AvatarBooth: high-quality and customizable 3D human avatar generation [EB/OL]. [2023-08-20]. https://arxiv.org/pdf/2306.09864.pdf

Zhang J R, Zhang Y S, Cun X D, Huang S L, Zhang Y, Zhao H W, Lu H T and Shen X. 2023a. T2M-GPT: generating human motion from textual descriptions with discrete representations [EB/OL]. [2023-08-20]. https://arxiv.org/pdf/2301.06052.pdf

Zhang L W, Qiu Q W, Lin H Y, Zhang Q X, Shi C, Yang W, Shi Y, Yang S B, Xu L and Yu J Y. 2023b. DreamFace: progressive generation of animatable 3D faces under text guidance. ACM Transactions on Graphics, 42(4): #138 [DOI: 10.1145/3592094]

Zhang M Y, Cai Z G, Pan L, Hong F Z, Guo X Y, Yang L and Liu Z W. 2022. MotionDiffuse: text-driven human motion generation with diffusion model [EB/OL]. [2023-08-20]. https://arxiv.org/pdf/2208.15001.pdf

Zheng Y F, Abrevaya V F, Bühler M C, Chen X, Black M J and Hilliges O. 2022. I M avatar: implicit morphable head avatars from videos//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 13535-13545 [DOI: 10.1109/CVPR52688.2022.01318]

Zhou H, Sun Y S, Wu W, Loy CC, Wang X G and Liu Z W. 2021. Pose-controllable talking face generation by implicitly modularized audio-visual representation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 4174-4184 [DOI: 10.1109/CVPR46437.2021.00416]

Zhou Y, Han X T, Shechtman E, Echevarria J, Kalogerakis E and Li D Z Y. 2020. MakeItTalk: speaker-aware talking-head animation. ACM Transactions on Graphics, 39(6): #221 [DOI: 10.1145/3414685.3417774]

Zhou Z W, Wang Z, Yao S Y, Yan Y C, Yang C, Zhai G T, Yan J C and Yang X K. 2022. DialogueNeRF: towards realistic avatar face-to-face conversation video generation [EB/OL]. [2023-08-20]. https://arxiv.org/pdf/2203.07931v1.pdf

作者简介

高玄，男，博士研究生，主要研究方向为高真实感数字人建模、多模态数字人驱动和隐式神经表示。
E-mail: gx2017@mail.ustc.edu.cn
张举勇，通信作者，男，教授，博士生导师，主要研究方向为计算机图形学、计算机视觉和数值优化。
E-mail: juyong@ustc.edu.cn
刘东宇，男，硕士研究生，主要研究方向为零样本语音合成。
E-mail: ld020519@mail.ustc.edu.cn