

中图法分类号: TP37 文献标识码: A 文章编号: 1006-8961(2025)02-0334-27

论文引用格式: Pan Y, Li S X, Tan S, Wei J J, Zhai G T and Yang X K. 2025. Advancements in digital character stylization, multimodal animation, and interaction. Journal of Image and Graphics, 30(02): 0334-0360 (潘焱, 李韶旭, 谭帅, 韦俊杰, 翟广涛, 杨小康. 2025. 数字人风格化、多模态驱动与交互进展. 中国图象图形学报, 30(02): 0334-0360) [DOI: 10. 11834/jig. 230639]

数字人风格化、多模态驱动与交互进展

潘焱*, 李韶旭, 谭帅, 韦俊杰, 翟广涛, 杨小康

上海交通大学计算机科学与工程系, 上海 200240

摘要: 风格化数字人是在计算机图形学、视觉艺术和游戏设计等领域中迅速发展的一个领域。数字人物的设计和制作技术取得了显著的进步, 使得数字人物能够具有更加逼真的外观和行为, 同时也可以更好地适应各种艺术风格和情境。本文围绕风格化数字人任务, 围绕数字人的风格化生成、多模态驱动与用户交互 3 个核心研究方向的发展现状、前沿动态、热点问题等进行系统性综述。针对数字人的风格化生成, 从显式三维模型和隐式三维模型两种数字人的三维表达方式对于方法进行分类。显式三维数字人风格化以基于优化的方法、基于生成对抗网络的方法、基于引擎的方法为主要分析对象; 隐式三维数字人风格化从通用隐式场景风格化方法以及针对人脸的隐式风格化进行回顾。针对数字人的驱动, 根据驱动源的不同, 从显式音频驱动、文本驱动和视频驱动 3 个方面进行回顾。根据驱动实现算法的不同, 从基于中间变量、基于编码—解码结构等方面进行回顾。此外, 算法还根据中间变量的不同可分为基于关键点、三维人脸和光流的方法。针对数字人的用户交互, 目前主流的交互方式是语音交互, 本文对语音交互模块从自动语音识别和文本转语音合成两方面进行了回顾, 对于数字人的对话系统模块, 从自然语言理解和自然语言生成等方面进行了回顾。在此基础上, 展望了风格化数字人研究的未来发展趋势, 为后续的相关研究提供参考。

关键词: 风格化; 数字人; 人脸驱动; 人机交互; 三维建模; 深度学习; 神经网络

Advancements in digital character stylization, multimodal animation, and interaction

Pan Ye*, Li Shaoxu, Tan Shuai, Wei Junjie, Zhai Guangtao, Yang Xiaokang

Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Abstract: Stylized digital characters have emerged as a fundamental force in reshaping the landscape of computer graphics, visual arts, and game design. Their unparalleled ability to mimic human appearance and behavior, coupled with their flexibility in adapting to a wide array of artistic styles and narrative frameworks, underscores their growing importance in crafting immersive and engaging digital experiences. This comprehensive exploration delves deeply into the complex world of stylized digital humans, explores their current development status, identifies the latest trends, and addresses the pressing challenges that lie ahead in three foundational research domains: the creation of stylized digital humans, multimodal driving mechanisms, and user interaction modalities. The first domain, creation of stylized digital humans, examines the

收稿日期: 2023-09-12; 修回日期: 2024-09-06; 预印本日期: 2024-09-13

* 通信作者: 潘焱 whitneypanye@sjtu.edu.cn

基金项目: 国家自然科学基金项目(62472285, 62102255); 上海市科技重大专项项目(2021SHZDZX0102)

Supported by: National Natural Science Foundation of China (62472285, 62102255); Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102)

innovative methodologies employed in generating lifelike but stylistically diverse characters that can seamlessly integrate into various digital environments. From advancements in 3D modeling and texturing to the integration of artificial intelligence for dynamic character development, this section provides a thorough analysis of the tools and technologies that are pushing the boundaries of what digital characters can achieve. In the realm of multimodal driving mechanisms, this study investigates evolving techniques for animating and controlling digital humans by using a range of inputs, such as voice, gesture, and real-time motion capture. This section delves into how these mechanisms not only enhance the realism of character interactions but also open new avenues for creators to involve users in interactive narratives in more meaningful ways. Finally, the discussion of user interaction modalities explores the various ways in which end-users can engage with and influence the behavior of digital humans. From immersive virtual and augmented reality experiences to interactive web and mobile platforms, this segment evaluates the effectiveness of different modalities in creating a two-way interaction that enriches the user's experience and deepens their connection to digital characters. At the heart of this exploration lies the creation of stylized digital humans, a field that has witnessed remarkable progress in recent years. The generation of these characters can be broadly classified into two categories: explicit 3D models and implicit 3D models. Explicit 3D digital human stylization encompasses a range of methodologies, including optimization-based approaches that meticulously refine digital meshes to conform to specific stylistic attributes. These techniques often involve iterative processes that adjust geometric details, textures, and lighting to achieve the desired aesthetic. Generative adversarial networks, as cornerstones of deep learning, have revolutionized this landscape by enabling the automatic generation of novel stylized forms that capture intricate nuances of various artistic styles. Furthermore, engine-based methods harness the power of advanced rendering engines to apply artistic filters and affect real time, offering unparalleled flexibility and control over the final visual output. Implicit 3D digital human stylization draws inspiration from the realm of implicit scene stylization, particularly via neural implicit representations. These approaches offer a more holistic and flexible approach for representing and manipulating 3D geometry and appearance, enabling stylization that transcends traditional mesh-based limitations. Within this framework, facial stylization holds a special place, requiring a profound understanding of facial anatomy, expression dynamics, and cultural nuances. Specialized methods have been developed to capture and manipulate facial features in a nuanced and artistic manner, fostering a level of realism and emotional expressiveness that is crucial for believable digital humans. Animating and controlling the behavior of stylized digital humans necessitates the use of diverse driving signals, which serve as the lifeblood of these virtual beings. This study delves into three primary sources of these signals: explicit audio drivers, text drivers, and video drivers. Audio drivers leverage speech recognition and prosody analysis to synchronize digital human movements with spoken language, enabling them to lip-sync and gesture in a natural and expressive manner. Conversely, text drivers rely on natural language processing (NLP) techniques to interpret textual commands or prompts and convert them into coherent actions, allowing for a more directive form of control. Video drivers, which are perhaps the most advanced in terms of realism, employ computer vision algorithms to track and mimic the movements of real-world actors, providing a seamless bridge between the virtual and physical worlds. These drivers are supported by sophisticated implementation algorithms, many of which rely on intermediate variable-driven coding-decoding structures. Keypoint-based methods play a pivotal role in capturing and transferring motion, allowing for the precise replication of movements across different characters. Moreover, 3D face-based approaches focus on facial animation and utilize detailed facial models and advanced animation techniques to achieve unparalleled realism in expressions and emotions. Meanwhile, optical flow-based techniques offer a holistic approach to motion estimation, synthesis, capture, and reproduction of complex motion patterns across the entire digital human body. The true magic of stylized digital humans lies in their ability to engage with users in meaningful and natural interactions. Voice interaction, currently the mainstream mode of communication, relies heavily on automatic speech recognition for accurate speech-to-text conversion and text-to-speech synthesis for generating natural-sounding synthetic speech. The dialog system module, a cornerstone of virtual human interaction, emphasizes the importance of natural language understanding for interpreting user inputs and natural language generation for crafting appropriate responses. When these capabilities are seamlessly integrated, stylized digital humans are capable of engaging in fluid and contextually relevant conversations with users, fostering a sense of intimacy and connection. The study of stylized digital characters will likely continue its ascendancy, fueled by advancements in deep learning, computer vision, and

NLP. Future research may delve into integrating multiple modalities for richer and more nuanced interactions, pushing the boundaries of what is possible in virtual human communication. Innovative stylization techniques that bridge the gap between reality and fiction will also be explored, enabling the creation of digital humans that are both fantastic and relatable. Moreover, the development of intelligent agents capable of autonomous creativity and learning will revolutionize the way stylized digital humans can contribute to various industries, including entertainment, education, healthcare, and beyond. As technology continues to evolve, stylized digital humans will undoubtedly play an increasingly substantial role in shaping how people engage with digital content and with each other, ushering in a new era of digital creativity and expression. This study serves as a valuable resource for researchers and practitioners alike, offering a comprehensive overview of the current state of the art and guiding the way forward in this dynamic, exciting field.

Key words: stylization; digital characters; face driven; human-computer interaction; 3D modeling; deep learning; neural network

0 引言

近年来,风格化数字人在视觉艺术和游戏设计等领域备受关注。这些具有独特风格和表现力的数字人物为创新性的艺术表达提供了无限可能性。本文对数字人的风格化生成、多模态驱动与用户交互等研究方向的发展现状、前沿动态、热点问题等进行系统性综述。

风格化生成旨在将不同的艺术风格应用于数字人物,以创造出具有独特外观和表现力的数字形象。多模态驱动是指通过多种模态的数据,如文本、语音、图像等,来驱动数字人物的言行和表现。多模态驱动能够增强数字人物的交互性和表现力,使其能够更自然地与用户进行交流和互动。用户交互是数字人研究的另一个重要方向。数字人需要能够与用户进行自然、流畅的交互,理解用户的需求和意图,并能够根据不同的场景和语境做出合适的回应。用户交互涉及到语音识别、自然语言处理、机器学习等领域,是实现智能交互的关键技术之一。

风格化生成作为数字人技术的起点,为数字人物注入了独特的艺术灵魂。这一过程不仅关乎外观的塑造,如面部特征、服饰风格、光影效果等,更深入到行为模式、性格特征及情感表达等层面。风格化生成确保了每个数字人都具有鲜明的个性和差异化特征,从而在众多应用场景中脱颖而出。这种个性化设计不仅提升了用户的视觉体验,更为后续的交互过程奠定了情感基础,使得用户能够更容易地与数字人建立情感连接。多模态驱动是数字人实现智能交互的关键环节。它允许数字人通过整合来自不同模态的信息(文本、语音、图像、视频、手势等),以更加全面和深入的方式理解用户的意图和需求。这

种多模态的感知与处理能力,使得数字人能够更加自然地融入用户的日常生活和工作环境中,实现更加流畅和高效的交互体验。多模态驱动不仅增强了数字人的表现力,还极大地提升了其交互的灵活性和适应性,为数字人应对复杂多变的交互场景提供了可能。用户交互是数字人技术的最终目标,也是衡量其成功与否的重要标准。它要求数字人能够与用户进行自然、流畅且富有情感交流,准确理解用户的意图和需求,并做出恰当的回应。这一过程涉及到语音识别、自然语言处理和机器学习等众多先进技术,它们共同构成了实现智能交互的基石。用户交互的顺畅与否,直接影响到用户对数字人的信任度和满意度,进而决定了数字人在各个领域的应用前景。风格化生成、多模态驱动与用户交互之间存在着紧密的逻辑关系。风格化生成成为数字人提供了独特的个性和差异化特征,这是吸引用户并建立情感连接的基础;多模态驱动则使得数字人能够更加全面和深入地理解用户,实现更加智能和自然的交互;而用户交互的顺畅与否,则直接反馈到风格化生成和多模态驱动的改进与优化中。这3个方向相互依存,共同推动着数字人技术的不断进化和完善。

风格化生成、多模态驱动与用户交互是数字人技术的3大支柱,它们之间的逻辑关系构成了数字人智能交互的完整框架。未来,随着技术的不断进步和应用场景的不断拓展,这3个方向的研究和发展将继续为智能交互、文化创意和娱乐产业等领域带来新的机遇和挑战。

1 风格化三维人脸生成

本节首先简述了图像及视频人脸风格化,然后

分析了当前主流的风格化三维人脸生成方法。图像及视频人脸风格化是三维人脸风格化的基础。三维人脸的表示可分为显式三维人脸和隐式三维人脸,前者一般使用网格模型(mesh model)作为三维表达方式,后者一般使用神经辐射场(neural radiance fields, NeRF)作为三维表达方式。显式三维人脸风格化方法主要包括基于优化的方法、基于生成对抗网络的方法。隐式三维人脸作为一种特殊的隐式三维场景,既适用于通用隐式场景风格化方法,也适用于针对人脸的隐式风格化。风格化目标的输入可以是手动编辑、文本引导、单幅图像、游戏引擎和图像数据集等。

1.1 图像及视频人脸风格化

1.1.1 图像风格化

图像风格化旨在实现真实图像与艺术风格之间的风格转换。从Gatys等人(2016)的工作开始,大量的相关工作(Cai等,2023)实现了高质量的面向通用图像的风格转换。然而,由于这些方法在保持结构完整性和细节保真度上存在局限性,很难直接应用于人脸风格化。近年来,生成模型的出现,为入脸到特定艺术风格的非线性映射提供了强有力的支持(廖远鸿等,2023)。在生成对抗网络(generative adversarial network, GAN)框架下,CycleGAN(Almahairi等,2018)、StyleGAN(Karras等,2019)等系列模型的提出为风格化人脸生成提供了高效、可控的解决方案。图1为风格化人脸生成展示,展示的风格

化目标的输入分别来自单幅图像(Han等,2023)、游戏引擎(Wang等,2023)、图像数据集(Wang等,2022a)和文本引导(Chen等,2023)。Stylegan-NADA(Gal等,2021)利用预训练的CLIP(contrastive language-image pretraining)模型来引导风格化生成器的微调训练,以实现跨域风格化图像的合成,如图2所示,其中 G (generator)表示生成器, W 为对抗神经网络的隐空间变量。

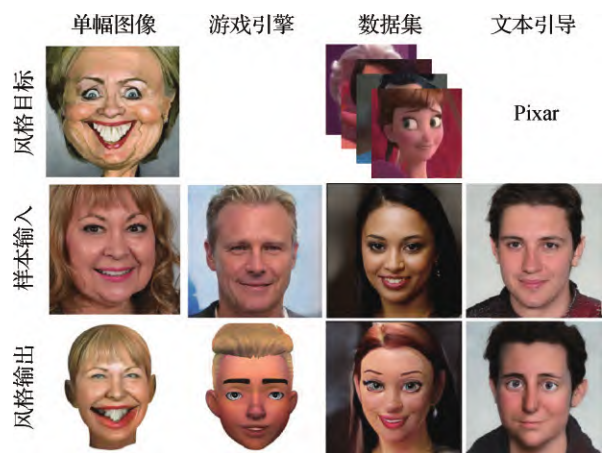


图1 风格化人脸生成展示

Fig. 1 Stylized face generation demonstration

最近,扩散模型(diffusion models)(Sohl-Dickstein等,2015)作为新兴的生成模型,在多模态、高质量上的表现优于生成对抗网络。相较于只能基于特定域图像训练的GAN,基于海量数据训练的扩散模型不

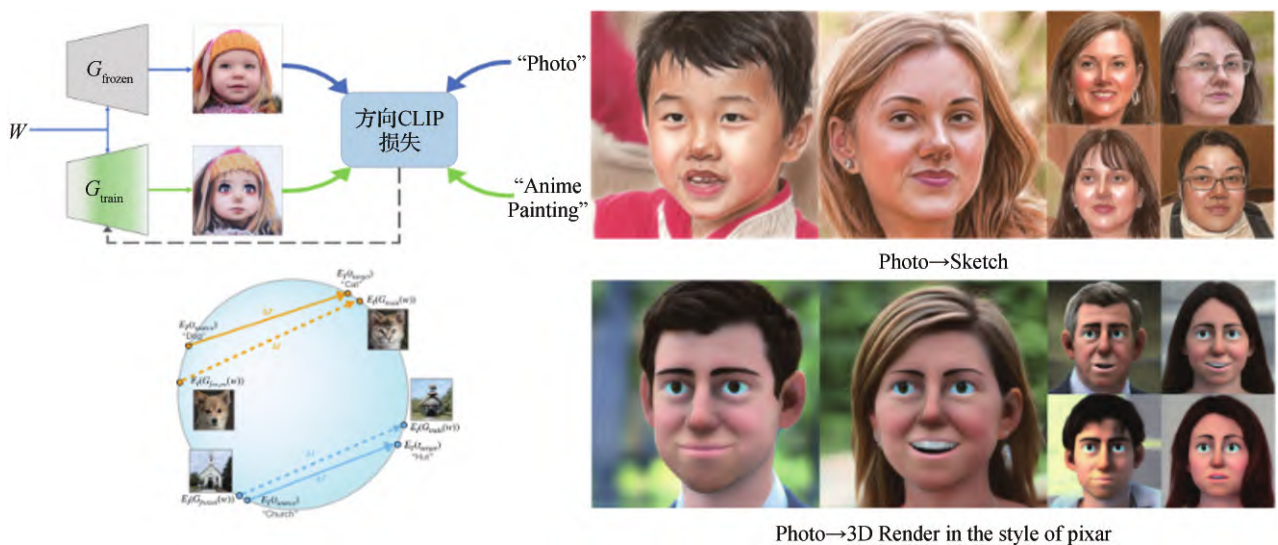


图2 Stylegan-NADA(Gal等,2021)利用预训练的CLIP模型引导风格化生成器的微调

Fig. 2 Stylegan-NADA(Gal et al., 2021) uses pre-trained CLIP models to guide fine-tuning of the stylization generator

受限于特定的图像域。潜在扩散模型(latent diffusion models)(Rombach等,2022)为在潜在空间编辑图像提供了一种便捷的方法。有了多重条件扩散模型(composable diffusion models)(Liu等,2022b),就出现了从多个输入创建图像的方法(刘安安等,2024)。例如,InstructPix2Pix(Brooks等,2023)可以通过图像输入和文本指令编辑图像。

1.1.2 视频风格化

与图像风格化类似,视频风格化旨在利用从风格图像中提取的风格特征再现给定视频的内容。与图像风格化相比,保持不同帧之间的一致性是一核心任务。针对这一任务,部分方法基于现有的图像风格化算法添加一致性损失函数来实现视频的风格化(Ruder等,2016;Chen等,2017;Gao等,2020)。近期,从文本到视频的合成引起了大量研究者的关注。如Video diffusion(Ho等,2022)使用基于时空因式分解的U-Net,联合图像和视频数据进行训练以实现视频生成。Rerender-A-Video(Yang等,2023b)使用自适应的扩散模型来生成关键帧,并应用分层跨帧约束来实现跨帧一致性,最后将关键帧传播到其他帧。Tune-A-Video(Wu等,2023)使用标准的扩散训练损失来更新注意力区块中的投影矩阵以实现对于文本到图像扩散模型的微调。与文本到图像相比,文本到视频由于缺乏高质量的视频数据集,其视频合成质量和一致性与商业应用相去甚远。

1.2 基于显式三维人脸的风格化

网格模型(mesh model)作为最广泛应用的三维表达方式,通常使用一系列多边形(通常是三角形)来近似表示三维物体的模型。作为一种简单易用的三维表达,mesh可以有效地完成三维人脸的静态和动态表达,在游戏、影视等领域有着广泛的应用(郝琮晖等,2024)。

1.2.1 基于优化的显式三维人脸的风格化

基于优化的方法首先重建真实三维人脸模型,再分别完成mesh的几何变形及纹理迁移,使其在保持原人脸身份特征的情况下,具有风格目标风格特征。Olivier等人(2022)提出用户控制的三维人脸漫画化系统。该系统由3部分组成:弯曲夸张模块控制人脸网格梯度的主成分分析(principal component analysis, PCA)分数来增强面部细节;比例夸张模块基于漫画三维人脸数据集生成大幅度的夸张形变;最终一个纹理和对比度模块负责调整生成三维

人脸的漫画细节,使其区别于常规人脸纹理。该系统为用户提供10个可调节3个模块参数的旋钮。MeinGame(Lin等,2021)通过形状迁移网络在维持拓扑不变的情况下将三维可变形模型(3D morphable model, 3DMM)(Blaiz和Vetter,1999)的人脸形状转换为游戏的人脸mesh。此外,该方法将输入图像展开至基于游戏mesh的UV(U和V为纹理贴图坐标的简称)空间得到粗纹理,通过编码解码后得到纹理特征。最后,形状、纹理以及光照系数被送入可微渲染器完成纹理及形状优化。Han等人(2023)提出一种基于样例的三维人脸风格化方法,给定一幅内容人脸图和一幅风格人脸图,该方法可生成具有夸张几何和纹理的风格化三维人脸模型。该方法的风格化框架由几何转换和纹理转换两部分组成。几何风格转换使用人脸关键点来引导常规三维人脸的几何形变,纹理转换阶段使用基于可微渲染的多视图优化完成纹理迁移。Li和Pan(2023)提出将重建的三维人脸从两个视角渲染为二维图像,借助于具有几何形变的图像风格化方法,以一致性损失作为约束对两视角图像同时风格化,再基于风格化后的图像完成风格化的三维人脸重建。此类方法得到的人脸通常能够满足几何形变的预期,但无法得到高质量的风格化人脸纹理。Face-to-Parameter(Shi等,2019)提出通过面部搜索来实现风格化人脸创建。该方法由模仿网络和特征提取器组成。模仿网络模拟游戏引擎的行为,即通过输入用户自定义的面部参数来生成游戏角色面部。特征提取器基于给定的人脸图像以及渲染的游戏人脸进行特征提取,并进行相似性度量和最优面部参数搜索。ClipFace(Aneja等,2023)提出了一种同时预测变形模型的纹理和表达潜码的神经网络。通过利用基于预训练CLIP模型的可微分渲染和损失,以自监督的方式进行训练,实现了文本引导的三维人脸面部纹理生成。DreamFace(Zhang等,2023c)从人脸几何参数空间内随机采样的候选项中选择最佳的粗略几何模型,然后通过隐式扩散模型(latent diffusion model, LDM)雕刻几何细节,使头部模型更符合文本提示,如图3所示。

1.2.2 基于生成对抗网络的显式三维人脸的风格化

生成对抗网络(GAN)通过生成器(generator)和判别器(discriminator)两个模块的互相博弈来完成



图3 DreamFace(Zhang等,2023c)实现纹理风格化

Fig. 3 Texture stylization by DreamFace (Zhang et al. , 2023c)

对于目标数据概率分布的学习和目标生成。基于大量同目标域图片训练的2D-GAN可实现高质量的图像生成。在无3D监督的情况下,2D-GAN也包含了部分3D信息。Wang等人(2022a)提出使用2D-GAN完成3D卡通人脸的生成和驱动。该方法在卡通数据集上微调预训练的StyleGAN(Karras等,2019)人脸模型,以共享隐空间编码来实现从人脸图像到共通人脸的转换。通过基于人脸模型GAN隐空间语义方向的编码调整,该方法可实现姿态、表情和光照的控制。

除了基于大量图像完成训练的GAN,部分方法自建三维风格数据集,以GAN来完成二维到三维的映射。3D-Magic-Mirror(Guo等,2019)首先从一组中性表情照片中利用多视图来获取高质量的3D中性表情人脸模型,并对其变形以获得一组blendshape。基于输入的多视图照片,该方法使用预训练的Cari-GeoGAN(Cao等,2018)模型对每一帧完成风格化,并组合优化得到融合的纹理图,如图4所示。3D-Magic-Mirror训练两个变分自编码器(variational auto-encoder, VAE)和一个CycleGAN(Almahairi等,2018)来实现几何形变,VAE网络分别学习从隐空间到三维人脸和三维漫画脸的生成,CycleGAN则学习两个隐空间的映射。3D-CariGAN(Ye等,2023b)构建了一个大规模的三维漫画人脸数据集,并基于此设计了三维漫画人脸统计模型。输入人脸照片,对应漫画三维人脸几何通过三维漫画人脸统计模型直接由输入的照片优化得到,纹理则通过基于常规三维人脸与图像的关键点对齐优化得到。将得到的常规纹理与漫画三维模型组成即可得到有纹理的三维漫

画人脸。



图4 3D-Magic-Mirror(Guo等,2019)实现夸张几何生成

Fig. 4 Exaggerated geometry generation by 3D-Magic-Mirror (Guo et al. , 2019)

SwiftAvatar(Wang等,2023)提出使用共享隐空间编码的双域生成器,以此为基础桥接avatar引擎与真人图像以实现从真人到avatar的生成。双域生成器由一个固定的现实生成器和一个迁移学习的角色生成器组成,可以基于给定编码生成相应的现实图像和虚拟图像。训练过程随机采样avatar引擎的生成编码渲染图像,基于双域生成器进行GAN反演可以得到对应真实人脸图像。基于此成对数据,SwiftAvatar可根据指定的avatar引擎的合成与人脸对应的avatar。基于引擎的方法得到的风格化三维人脸受限于引擎,虽然能够得到较高的合成质量,但很难实现与真实人脸匹配的自由的风格化结果。

1.3 基于隐式三维人脸的风格化

神经辐射场(NeRF)作为最新的三维场景表示方法,可以在给定一组多视图图像的情况下实现高质量的新视角图像合成。人像作为特殊的三维场景,除了作为一般的静态NeRF场景(Jiang等,2022)表示外,还有部分研究借助可变形的NeRF(Zielonka等,2023)实现动态avatar合成。得益于NeRF高质量的特性,基于NeRF的avatar在渲染质量上优于传统的三维模型。

1.3.1 通用隐式场景风格化

NeRF-ART(Wang等,2024)提出了一种使用文

本引导 NeRF 风格化的方法,该方法基于训练好的 NeRF 场景,以 CLIP 来对风格化进行监督,并设计了一种基于剪辑的对比损失来适当地加强风格化,如图 5 所示。为了进一步确保整个场景风格的一致性,该方法将对对比约束扩展到全局—局部混合框架,以涵盖全局结构和局部细节。此外,为了支持与外观相结合的几何风格化,其放松了对预先训练的神经网络的密度的约束,并采用了一种权重正则化方法。



图5 NeRF-ART(Wang等,2024)实现静态通用
NeRF 场景风格化

Fig. 5 Static generic NeRF scene stylization by NeRF-ART
(Wang et al., 2024)

Instruct-NeRF2NeRF(Haque等,2023)提出使用自然语言指令编辑 NeRF 场景的方法。该方法使用条件扩散模型 InstructPix2Pix(Brooks等,2023)来对场景图像进行迭代编辑。该方法的核心是交替进行图像编辑与 NeRF 训练,在训练 NeRF 的过程中,场景被渲染至图像并由扩散模型进行更新,随后继续监督 NeRF 的训练。每次迭代时更新若干图像并训练 NeRF 若干循环次数。随着迭代的进行,扩散模型对 NeRF 图像的编辑逐渐收敛,实现预期的 NeRF 场景编辑。通用隐式场景风格化方法,可以实现对静态人像 NeRF 场景的风格化编辑,但此类方法无法完成高质量的动态 avatar 的风格化编辑。

1.3.2 针对人脸的隐式风格化

基于 NeRF 的高真实感 avatar 合成已有大量研究,现有方法可以在给定一段人物视频的情况下完成高质量的动态 avatar 合成与驱动。Instruct-Video2Avatar(Li,2023)使用 INSTA(instant volumetric head avatars)(Zielonka等,2023)作为 avatar 的表征方式,对于 Instruct-NeRF2NeRF 的编辑方法进行了优化,实现了对动态 avatar 的高质量的时间、空间一致性编辑。不同于 Instruct-NeRF2NeRF 对于场景所有图像使用扩散模型编辑,Instruct-Video2Avatar 使用扩散模型编辑一帧样例图像,使用视频风格化

方法编辑其他的视频帧,最后使用编辑后的视频合成风格化的 avatar。经过数次视频风格化及 avatar 训练迭代,可以得到高质量的风格化 avatar,如图 6 所示。



图6 Instruct-Video2Avatar(Li,2023)实现动态 avatar 风格化
Fig. 6 Dynamic avatar stylization by Instruct-Video2Avatar
(Li, 2023)

类似地,Instruct-NeuralTalker(Sun等,2023)同样使用迭代式方法来更新 avatar 的训练视频,但该方法使用不同的 avatar 表征方式且不采用视频风格化来辅助编辑。为了提升 avatar 的编辑效果,该方法从使用较小的编辑幅度开始,渐进式地增加编辑的修改幅度。此外,该方法使用嘴唇边界损失作为约束,以提升唇形的编辑一致性。

AlteredAvatar(Nguyen-Phuoc等,2023)以 Instant Avatar(Cao等,2022)作为 avatar 的表征方法,Instant Avatar 有身份和表情两个编码器,身份编码器将人物的几何和纹理使用 2D 卷积偏差映射器映射至多尺度偏差特征,并以此实现 mesh 的解码重建。AlteredAvatar 风格化过程使用 CLIP 来计算 avatar 渲染图像和文本或图像目标风格化损失,通过修改偏差映射器的网络权重,保持其余网络参数不变可实现风格化。为了快速风格化,该方法使用元学习(meta-learning)框架使用双循环来完成快速基于提速 avatar 风格迁移,在外循环中,模型学习优化多个目标样式,在内循环中,模型学习优化一个目标样式。Control4D(Shao等,2023a)提出高保真的时间一致性 4D 肖像编辑方法。该方法使用鉴别器来学习编辑图像的生成分布,并以此来更新生成器。具体地,该方法首先利用 Tensor4D(Shao等,2023b)来训练四维

肖像场景的隐式表示,然后使用体素渲染将其渲染为潜在特征和RGB图像,作为四维GAN的输入。该方法使用ControlNet(Zhang等,2023b)来完成渲染图像的编辑,以此作为GAN的判别器的真实输入来完成GAN的生成器和训练器的迭代更新。

1.3.3 基于三维生成对抗网络的风格化

基于大量数据训练的2D-GAN可以完成图像视角的调整,但其无法完全解耦图像特征和角度特征。3D-GAN(Chan等,2022)在训练时借助于从图像中提取的相机参数,在现实数据集上训练的3D-GAN通过引入隐式表达作为中间特征,可以实现高三维一致性的几何纹理生成。而艺术数据由于其丰富的风格化信息,阻碍了直接提取相机信息并完成3D-GAN的训练。3DAvatarGAN(Abdal等,2023)提出一

种将预训练的3D-GAN迁移至2D-GAN的目标域的方法。该框架首先基于优化完成跨域的相机参数对齐,并以纹理、深度和几何正则化作为约束,微调经过良好预训练的EG3D(efficient geometry-aware 3D generative adversarial networks)真实人脸网络模型。此外,该方法使用薄板样条的3D变形模块,来完成三维空间的大几何形变,如图7所示。HyperStyle3D(Chen等,2023)通过构建超网络来学习在前向传递中操纵生成器的参数。超网络将形状、属性和风格3个层次的文本编码为粗、中、细粒度的方向特征,并通过预训练的3D生成器预测3个层次的参数偏移。通过以CLIP和ID(identity)损失引导训练,该方法可以处理形状、属性和风格3个属性叠加下的文本引导的三维人脸风格化。

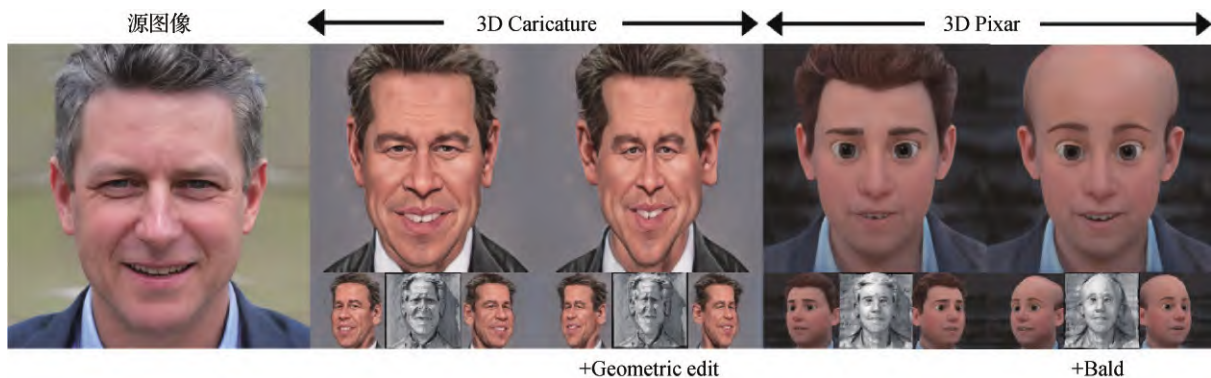


图7 3DAvatarGAN(Abdal等,2023)合成风格化人脸

Fig. 7 Stylization face synthesis by 3DAvatarGAN(Abdal et al., 2023)

1.4 对比与分析

风格化三维人脸生成的各种方法根据算法实现的不同,在各应用方面各有优劣。现有的游戏、影视等主流工作均使用显式模型,隐式模型的广泛应用有待行业相关技术的发展。风格化目标的输入可以是手动编辑、文本引导、单幅图像、游戏引擎和图像数据集。其中,手动编辑、文本引导、单幅图像不受限于特定风格,而基于游戏引擎和基于图像数据集的方法则受限于预设的数据风格。多数风格化生成方法实现了纹理和几何的同时编辑,部分显式三维人脸编辑只针对几何或纹理进行编辑。对于仅有纹理编辑的方法,三维人脸的驱动易于实现。对于有几何形变的显式三维模型,基于游戏引擎的方法可以驱动。采用文本引导的隐式风格化方法均可实现面向任意风格的纹理和几何编辑,但只有针对avatar的风格化才可实现动态驱动。

为了便于理解,图8给出了时间线用于表示各类工作的出现时间。表1为风格化人脸生成方法的总结。

2 多模态驱动数字人对话生成

多模态驱动数字人对话生成是指通过输入多种模态(音频、文本和视频)来驱动数字人进行对话的任务。目前,高质量的面部动画有着广泛的应用,例如,在教育领域中,高质量的面部动画可以用于生成虚拟教师以及虚拟学生,通过创造一个理想的学习环境来激起学生学习的动力,以此达到更好的教学效果;在幸福和健康领域中,所生成的数字人物不仅可以充当用于治疗聊天机器人,甚至还可以生成已故之人的影像,用于慰藉丧失至亲的悲伤情绪;在娱乐领域中,该技术更是在虚拟现实、电影制作等场

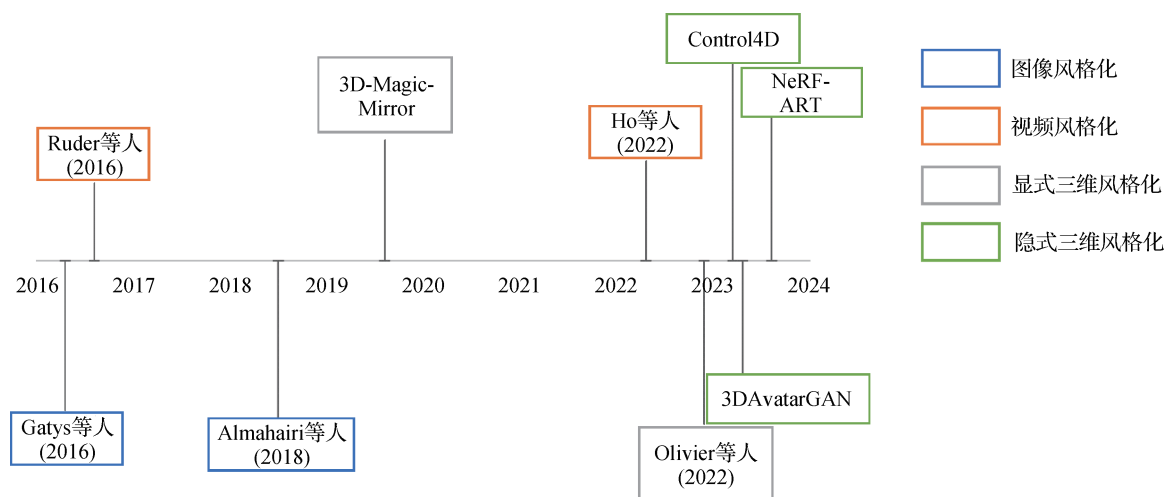


图8 风格化人脸生成方法的时间线

Fig. 8 Timeline of face stylization generation methods

景下发挥了极大的作用。然而,由于传统的动画制作方法往往需要大量人力物力的投入,如何自动驱动数字人进行自然的对话交流仍然是一项具有挑战性的任务。

随着人民生活的逐步改善以及神经网络技术的不断发展,近年来,虚拟现实等娱乐方式逐渐普及,同时越来越多的学者也投身于将深度学习技术应用到虚拟现实、增强现实等研究领域中来,当前主流的多模态驱动方法可分为显式音频驱动,文本驱动和视频驱动。图9为多模态驱动数字人对话生成展示。为了便于理解,图10给出了时间线用于表示各类工作的出现时间。

2.1 音频驱动的对话人脸生成

音频驱动的对话人脸生成指通过一段音频驱动一幅静态的图像进行对话的任务,当前的方法大致可以分为基于中间变量表示的生成方法和基于编码—解码结构的重建生成方法。由于该任务中输入的音频和输出的图像模态不同,增加了该生成任务的难度。基于中间变量表示的生成算法通常利用中间表示,如标志点、3D模型参数以及光流运动场等,作为链接两模态间的桥梁。Chen等人(2019)首先从音频和初始图像的标志点预测出与音频同步的标志点序列,随后利用生成的标志点和初始图像生成相应的视频帧序列,构成输出视频。但他们的结果中音频和嘴型的同步性较差,为解决这一问题,Das等人(2020)在生成标志点位置时引入了生成对抗网络(GAN),利用GAN学习因音频而产生的嘴型的微妙变形的强大性能,该方法有效地提升了嘴型的准确

性。然而,上述方法中主要预测嘴型和音频的同步,忽略了一个影响生成视频的真实性的因素:头部姿态。因此,Zhou等人(2020)将音频分为内容编码和身份编码,分别用于预测嘴部区域的标志点和头部姿态的标志点,最后生成具有个性化头部姿态的对话人脸视频,如图11和图12所示。

利用3D模型参数作为中间变量的方法一般是引入3DMM对输入的人脸图像进行重建得到相应的参数,然后对参数进行预测。从音频中预测表情参数,重建下半脸的图像后与输入图像融合,实现了对话人脸视频的生成。Song等人(2022b)首先从输入的图像中获取3D参数,其中表情参数用从音频中预测到的表情参数替换,然后进行人脸图像的重建。同样地,这些方法虽然实现了嘴型和音频的同步,但也忽略了对头部姿态的建模。Yi等人(2020)从音频中额外预测了姿态参数用于预测头部姿态的变化,由于轻量级图形引擎(Genova等,2018)渲染的合成帧真实性欠佳,作者引入了Memory Network对图像质量进行优化。在此基础上,Zhang等人(2021)综合考虑了局部特征、全局特征与嘴型、头部姿态和眨眼的相关性,实现了自然的头部运动、眨眼和嘴型的预测。

基于编码—解码结构的重建生成方法通常将输入的音频和图像/视频利用编码器映射到高维空间中进行融合,然后输入进解码器合成音频嘴型同步的视频。Jamaludin等人(2019)为保留身份特征给图像编码器输入多幅图像,并采用跳跃连接传输特征给解码器,但没有考虑帧之间的连续性,导致视频

表1 风格化人脸生成方法总结
Table 1 Summary of face stylization generation methods

	类别	方法	风格化目标输入	任意风格	编辑类型	动态驱动
图像风格化	基于优化的风格化	Gatys 等人(2016)	单幅图像	√	纹理	×
	基于生成对抗网络的风格化	Almahairi 等人(2018)	图像数据集	√	纹理	×
		Karras 等人(2019)	图像数据集	√	纹理+几何	×
		Gal 等人(2021)	文本引导	√	纹理+几何	×
	基于扩散模型的风格化	Sohl-Dickstein 等人(2015)	文本引导	√	纹理+几何	×
		Rombach 等人(2022)	文本引导	√	纹理+几何	×
Liu 等人(2022b)		文本引导	√	纹理+几何	×	
视频风格化	基于预训练网络的风格化	Ruder 等人(2016)	单幅图像	√	纹理	×
		Chen 等人(2017)	单幅图像	√	纹理	×
		Gao 等人(2020)	单幅图像	√	纹理	×
	基于扩散模型的风格化	Ho 等人(2022)	文本引导	√	纹理+几何	×
		Wu 等人(2023)	文本引导	√	纹理+几何	×
		Yang 等人(2023b)	文本引导	√	纹理	×
显式三维人脸	基于优化的风格化	Olivier 等人(2022)	手动引导	√	几何	×
		MeinGame(Lin 等,2021)	游戏引擎	×	纹理+几何	√
		Face-to-Parameter(Shi 等,2019)	游戏引擎	×	纹理+几何	×
		Han 等人(2023)	单幅图像	√	纹理+几何	×
		Li 和 Pan(2023)	单幅图像	√	纹理+几何	×
		ClipFace(Aneja 等,2023)	文本引导	√	纹理	√
		DreamFace(Zhang 等,2023c)	文本引导	√	纹理	√
	基于生成对抗网络的风格化	3D-Magic-Mirror(Guo 等,2019)	图像数据集	×	纹理+几何	×
		SwiftAvatar(Wang 等,2023)	游戏引擎	×	纹理+几何	√
		隐式三维人脸	通用隐式场景风格化	NeRF-ART(Wang 等,2024)	文本引导	√
Instruct-NeRF2NeRF(Haque 等,2023)	文本引导			√	纹理+几何	×
针对 avatar 的隐式风格化	Instruct-Video2Avatar(Li,2023)		文本引导	√	纹理+几何	√
	Instruct-NeuralTalker(Sun 等,2023)		文本引导	√	纹理+几何	√
	AlteredAvatar(Nguyen-Phuoc 等,2023)		文本引导	√	纹理+几何	√
	Control4D(Shao 等,2023a)		文本引导	√	纹理+几何	√
基于三维生成对抗网络的风格化	3DAvatarGAN(Abdal 等,2023)		图像数据集	×	纹理+几何	×
	HyperStyle3D(Chen 等,2023)	图像数据集	×	纹理+几何	×	

注:“√”表示可实现,“×”表示不可实现。

帧之间的抖动。Song 等人(2019)引入循环神经网络(recurrent neural network, RNN)(Zaremba 等, 2015), 考虑了音频的时序性, 有效地解决了帧间的抖动问题, 但仍存在嘴型和音频不同步的问题。Prajwal 等人(2020)利用一个预训练的嘴型同步的判别器来辨别生成的嘴型的同步性, 使得生成器生成与音频更加同步的嘴型。然而, 上述方法中没有考虑到音频和视频耦合的音频相关和说话者相关的特征, 而

是直接从音频和图像中预测结果。因此, Zhou 等人(2019)通过关联和对抗训练, 从视频输入中分离出语音相关特征和身份相关特征。Zhou 等人(2021)进一步从视频输入中分解出身份空间、姿势空间和语音内容空间, 实现了姿态的控制。除嘴型和头部姿态外, 表情也是影响生成结果真实性的关键因素。近年来, 带有情感的对话人脸生成方法受到了越来越多的关注, Karras 等人(2017)学习了一个情感隐



图9 多模态驱动数字人对话生成展示

Fig. 9 Demonstration of multimodal driven talking head faces generation

性空间,并在测试时控制情感种类,生成动态表情。但由于数据有限,他们的方法不能覆盖到所有的情绪。最近,Wang等人(2020)公开了一个大型的带有情感的视频音频数据集 MEAD(multi-view emotional

audio-visual dataset),并使用 one-hot 向量来编码情感。类似地,Eskimez 等人(2022)通过一幅图像和独热编码(one-hot)情感向量从音频中生成带有情感的对话人脸视频。然而,音频中蕴含的情感信息较少,不能用于合成精细的表情(陶建华等,2024)。因此, Ji 等人(2022)将遮挡嘴部的情感视频作为表情来源,但他们提取出的情感信息主要表示局部的情感偏移,忽略了情感对其他因素的影响,如嘴部等。

上述方法主要讨论二维(2D)对话人脸的生成工作,除此之外,三维(3D)对话人脸的生成同样也是受到学者欢迎的研究领域。早期的方法(Edwards等,2016;Ezzat和Poggio,2002;Kalberer和Van Gool,2002;Verma等,2003)主要是利用程序规则对预定义的面部参数进行动画制作。基于隐马尔可夫模型(hidden Markov model, HMM)的模型可从输入文本

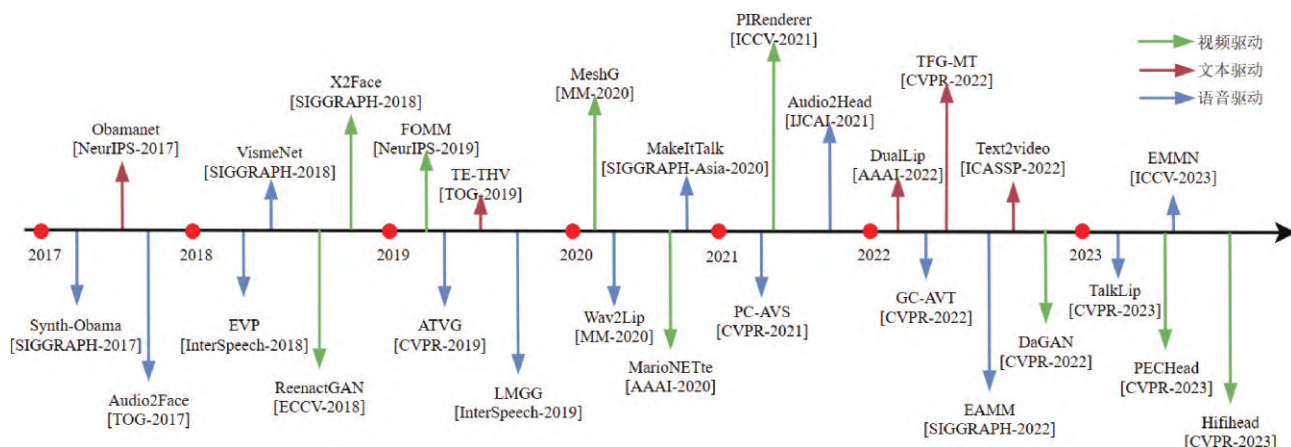


图10 多模态驱动数字人对话生成方法的时间线

Fig. 10 Timeline of the multimodal driven digital human dialog generation approach

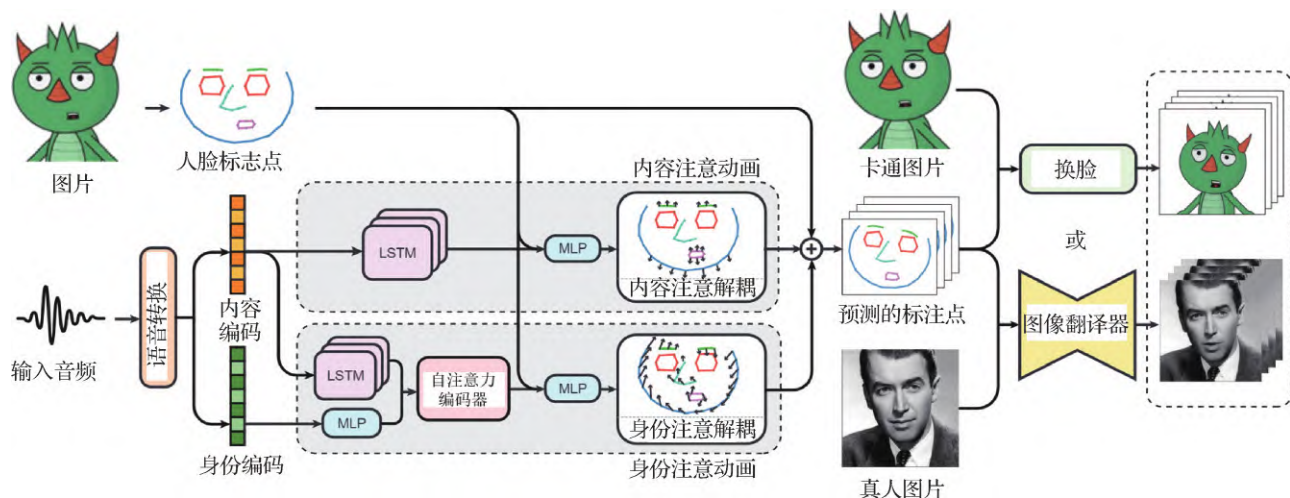


图11 MakeItTalk(Zhou等,2020)利用语音驱动关键点来合成人脸

Fig. 11 MakeItTalk utilizes speech-driven keypoints to synthesize faces(Zhou et al., 2020)

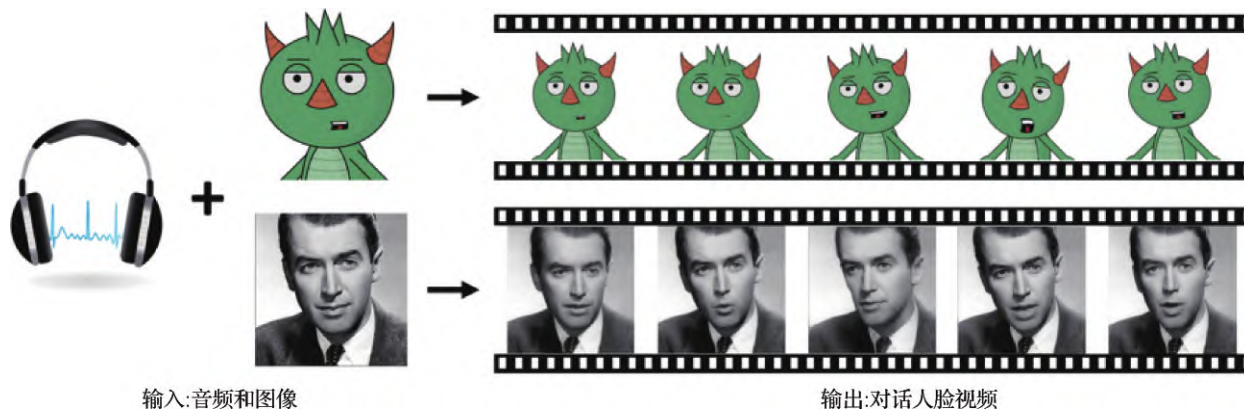


图12 MakeItTalk(Zhou等,2020)生成的结果

Fig. 12 The results generated by MakeItTalk(Zhou et al. , 2020)

或音频生成视觉效果,而面部动画则是通过视觉效果相关的共同发音模型(Edwards等,2016)或混合面部模板(Kalberer和Van Gool,2002)生成的。随着机器学习技术的不断进步,数据驱动的方法(Cao等,2005;Cudeiro等,2019;Fan等,2022)已经证明了其从数据中学习 viseme 模式的能力。这些方法基于预训练的语音模型(Baevski等,2020;Schneider等,2019)提取输入音频的特征表示,然后由卷积神经网络或自回归模型进行编码,以映射到3DMM空间或直接映射到3D网格。Audio2Face(Tian等,2019)从3~5 min的高质量说话者特定三维数据中学习三维面部动画模型,如图13所示。VOCA(voice operated character animation)(Sak等,2014)是在多个说话者的三维数据上进行训练的,通过在推理过程中提供一个 one-hot 形式的单次编码指定说话者,可以从输入音频中生成相应的身份动画。MeshTalk(Richard等,2021)是一种通用方法,它可以学习面部表情的

分类表示,并从该分类空间自动递归采样,从而根据音频输入将给定的说话者三维面部网格制作成动画。FaceFormer(Fan等,2022)使用预训练的 Wav2Vec(Schneider等,2019)音频特征提取器,并应用基于Transformer的解码器预测最终结果。与VOCA一样,FaceFormer 也为解码器提供了一个说话者识别码,允许用户从训练集中选择说话风格。

值得注意的是,近年来,神经辐射场(NeRF)(Mildenhall等,2021)被引入到音频驱动的人像合成中。它提供了一种新的方法,通过深度多层感知器(multi-layer perceptron,MLP)学习从音频特征到相应视觉外观的直接映射。此后,一些研究以端到端方式(Guo等,2021;Yao等,2022;Liu等,2022c;Shen等,2022)或通过一些中间表征(Ye等,2023a;Chatziagapi等,2023)对音频信号进行NeRF调节,以重建特定的对话人脸生成。具体来说,早期的工作主要建立在虚构的NeRF渲染器上。其中,AD-NeRF

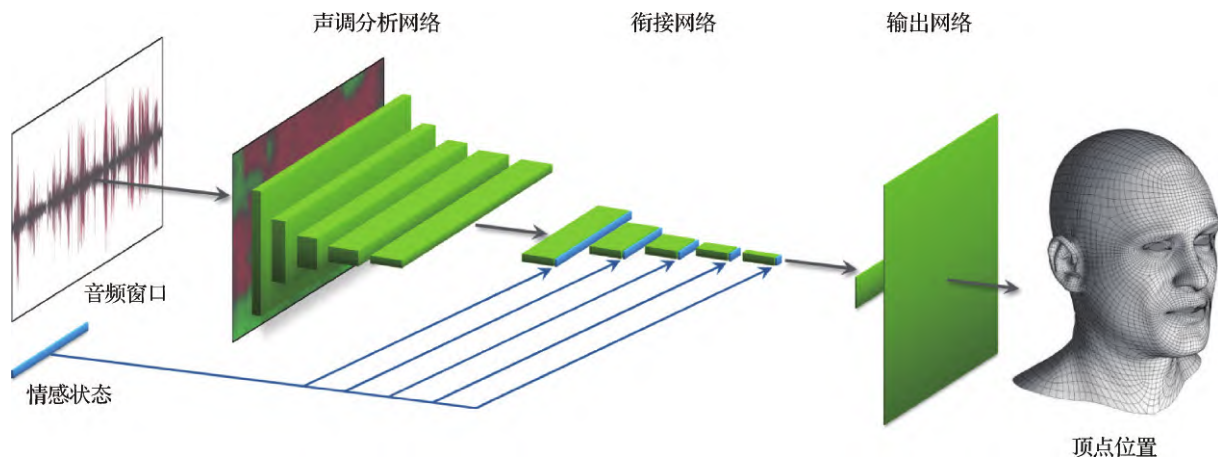


图13 Audio2Face(Tian等,2019)网络结构

Fig. 13 Network architecture of Audio2Face(Tian et al. , 2019)

(audio driven neural radiance fields)(Guo等,2021)首次将NeRF应用到对话人脸视频生成领域,并实现了高清的人脸渲染。SSP-NeRF(semantic-aware implicit neural radiance fields)(Liu等,2022c)考虑了音频对面部区域的不同影响,并采用语义采样策略鼓励局部运动建模。RAD-NeRF(real-time neural radiance)(Tang等,2022)以实时视频生成为目标,采用了基于Instant-NGP(instant neural graphics)(Müller等,2022)的NeRF。不过,它需要一个复杂的模块来处理音频信号。这些端到端方法将整个或部分大型MLP网络作为编码器来学习音频和区域

之间的连接,从而增加了其复杂性和训练难度。一些多阶段方法(Ye等,2023b;Chatziagapi等,2023)预先训练一个模型,通过中间表征学习视听关系,并利用基于NeRF的渲染器生成图像,例如GeneFace(Ye等,2023b)尝试通过将语音特征转化为面部地标来减少NeRF伪影,但这往往会导致唇部动作不准确。ER-NeRF(efficient region-aware neural radiance fields)(Li等,2023)创新性地引入了三平面哈希编码器来修剪空的空间区域,倡导一种紧凑、加速的渲染方法。

表2为音频驱动的对话人脸生成方法的总结。

表 2 音频驱动的对话人脸生成方法总结
Table 2 Summary of audio driven talking face generation methods

类别	方法	输入	任意人脸	头部姿态	表情
标志点	Chen等人(2019)	图像+音频	√	×	×
	Das等人(2020)	图像+音频	√	×	×
	Zhou等人(2020)	图像+音频	√	√	×
	Wang等人(2020)	图像+音频	×	×	√
	Ji等人(2021)	图像/视频+音频	×	×	√
基于中间变量表示	Karras等人(2017)	音频	×	×	√
	Thies等人(2020)	图像/视频+音频	×	×	×
	Song等人(2022b)	图像+音频	√	×	×
	Yi等人(2020)	图像/视频+音频	√	√	×
	3D模型参数	Zhang等人(2021)	√	√	×
	Wen等人(2020)	图像+音频	√	×	×
	Zhang等人(2022c)	图像+音频	√	×	×
	Cudeiro等人(2019)	音频	√	×	×
	Zhang等(2023d)	图像+音频	√	√	×
	光流运动场	Wang等人(2021)	√	√	×
	Wang等人(2022b)	图像+音频	√	√	×
	Ji等人(2022)	图像+视频+音频	√	√	√
	Yin等人(2022)	图像+音频	√	×	×
基于编码—解码结构	Jamaludin等人(2019)	图像+音频	√	×	×
	Chen等人(2018)	图像+音频	√	×	×
	Sadoughi和Busso(2019)	图像+音频	√	×	√
	Song等人(2019)	图像+音频	√	×	×
	Prajwal等人(2020)	图像/视频+音频	√	√	×
	Zhou等人(2019)	图像/视频+音频	√	×	×
	Zhou等人(2021)	图像/视频+音频	√	√	×
	Eskimez等人(2022)	图像+音频	√	×	√
	Park等人(2022)	图像+音频	√	×	×

注:“√”表示可实现,“×”表示不可实现。

2.2 文本驱动的对话人脸生成

文本驱动的对话人脸生成任务主要旨在生成与文本内容相匹配的脸部视频,可以看做是音频驱动的对话人脸生成任务的扩展。因此,文本驱动对话人脸生成的一个重要分支的基本思想是利用文本转语音(text-to-speech, TTS)系统从文本中生成语音,再利用音频驱动的方法生成人脸视频,而另一个重要分支则直接从文本中提取特征,以端到端的结构直接生成对应的视频,避免了两阶段的生成过程。对于前者来说,Kumar等人(2017)首先从文本合成音频,再预测与音频同步的关键点,最后结合输入的图像生成最终的视频。但该方法在训练一个人时需要17小时的数据。Zhang等人(2022a)除了从文本中合成音频外还建立了一个音素—姿态字典,用于关键帧的生成,降低了对数据量的需求。类似地,Hu等人(2021)从文本中提取关键语义标签,作为表情和身体姿态的驱动信号。然而,上述方法都只针对特定语言(英语或汉语),不具有泛化性能。为解决此问题,Prajwal等人(2019)引入了文本到文本的翻译系统,再利用翻译得到文本合成对应语言的音频,实现了A语言讲话视频到同一个人B语言讲话视频的自动翻译过程,并改成原本视频的嘴型使其与B语言音频对齐,但仍只能实现指定的两种语言间翻译。Song等人(2022a)通过向系统中添加语言嵌入(language embedding)和演讲者嵌入(speaker embedding),实现了多种语言的音频合成以及相应的对话人脸视频合成。

基于端到端的文本驱动的对话人脸生成方法主要从文本中提取如音素、词向量等特征后通过编码—解码结构进行最终结果的生成。Yu等人(2019)通过TTS(Wu等,2016)提取二元语言学特征后输入Time-delay LSTM(time-delay long short term memory network)预测嘴部的标志点,最后与输入的视频结合生成最终的视频。而Fried等人(2019)则是从文本中提取音素信息以及对应的表情参数并记录,再通过不同音素和参数的组合实现了对输入视频中的文本和嘴型进行编辑的目的。考虑到嘴型主要与文本中的音素信息相关,而表情和头部姿态主要与文本的语义信息相关,Li等人(2021)从文本中提取音素和词向量,分别用于嘴型和表情、头部姿态的预测。但上述方法中大多采用自回归(auto-regressive, AR)解码方式,以之前生成的帧为条件生成当前的帧,减慢了预测的速度,且错误累计问题也会导致生成帧的质量下滑,因此,Liu等人(2022a)预测编码的语言特征的持续时间,以非自回归的方式对以编码的语言特征为条件,对目标帧进行建模,实现了高效的预测性能,如图14和图15所示。

2.3 视频驱动的对话人脸生成

视频驱动的对话人脸生成是驱动数字人按照给定的视频进行对话驱动,该任务在视频会议、多人在线游戏等领域中存在诸多应用。当前的视频驱动的对话人脸生成主要分为基于2D的驱动方法和基于3D的驱动方法。

在基于2D驱动的方法中,Wiles等人(2018)通

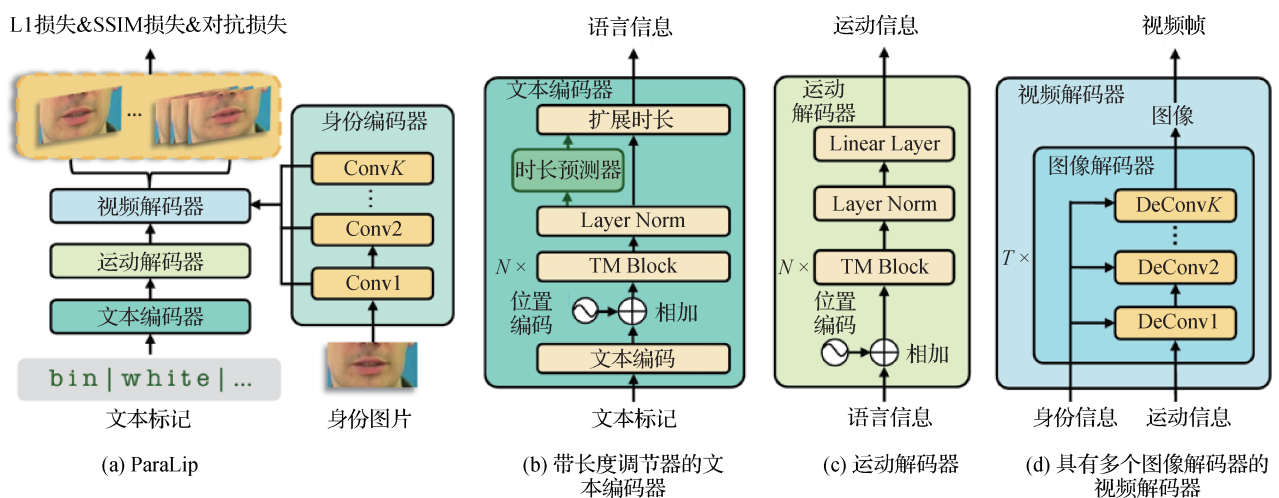


图14 ParaLip(Liu等,2022a)网络结构

Fig. 14 Network architecture of ParaLip(Liu et al., 2022a)((a) ParaLip; (b) text encoder with length regulator; (c) motion decoder; (d) video decoder with multiple image decoders)

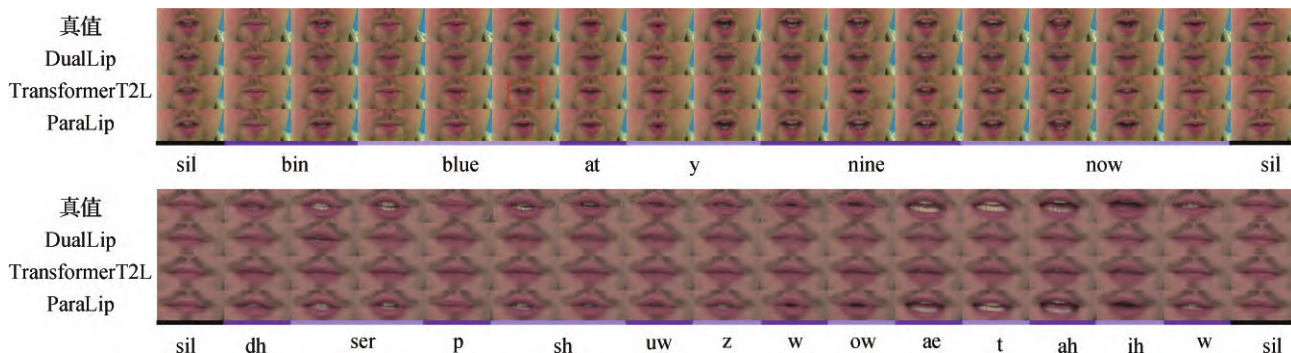


图 15 ParaLip(Liu 等, 2022a)的结果

Fig. 15 Results of ParaLip(Liu et al., 2022a)

过从驱动视频中提取到的运动表征对输入的人脸进行扭曲变形得到最终结果,但该方法生成的结果中原图像的身份特征损失较为严重。随后 Ha 等人(2020)设计了一个注意力模块用于更加有效地提取身份特征,增强输出结果的身份信息保留能力。Siarohin 等人(2020)提出光流评估模块,以无监督的方式从输入的图像和驱动视频中提取出若干关键点,并用这些关键点计算这两帧之间的光流运动场,进一步扭曲输入图像以生成与驱动视频一样的动作,但同时保留输入图像的身份和纹理。另外有一些方法利用 2D 标志点进行驱动生成。Averbuch-

Elor 等人(2017)和 Geng 等人(2018)先利用提取到的脸部标志点驱动输入的人脸生成一个粗糙的结果,随后再采用一个生成对抗网络对该结果进行精细化。类似地,Ha 等人(2020)提出的核心观点都是利用 68/98 个标志点表示驱动视频的运动,而进一步扭曲输入的图像。但这些方法所利用的标志点都较为稀疏,这将会导致生成的驱动流并不足够准确,进一步影响了生成的视频的质量。为解决这一问题,Zhang 等人(2023a)利用更加密集的 669 个头部标志点(Wood 等,2022)去生成更加丰富的头部几何信息以提升结果的表现性能,如图 16 所示。



图 16 MetaPortrait(Zhang 等, 2023a)语音驱动说话人脸照片

Fig. 16 Voice-driven talking face photo by MetaPortrait(Zhang et al., 2023a)

不同于基于 2D 的方法,3D 方法往往依赖 3D 模型(Blanz 和 Vetter, 2023)将人脸表征为与身份、表情和姿势相关的 PCA 系数,以此为编辑和渲染人像图像提供一个便利的操作方式。一个最常见的流程(Kim 等, 2019)是将源图像的身份系数与驱动视频

的运动系数进行组合,再通过渲染方法生成数字人驱动的视频。尽管这类方法允许用户进行显式的人脸控制以便捷地生成驱动结果,但由于 3D 模型参数的表征也较为稀疏,同时并不能对牙齿等部分进行建模,所以生成具有微表情且带有面部细节信息的

高分辨率视频仍然是当前方法的主要挑战。

2.4 对比与分析

当前多模态驱动数字人对话生成发展迅速,以不同模态作为输入的方法之间的比较可以从方法角度以及应用角度来进行分析和对比。从方法的角度来说,视频驱动的对话人脸生成发展和性能往往优于音频驱动的对话生成方法,这主要是由于视频中通常包含更多的有效信息,如直观的嘴型运动、头部运动等;而音频驱动的方法需要对从音频到对话人脸进行跨模态建模的过程,较为困难。而从应用的角度来说,音频驱动的方法所需要的条件较为简单,仅需要一段音频,而视频驱动的方法则需要一段完整的视频。在实际应用中也看重推理的速度,视频驱动的方法在对输入视频逐帧处理的速度通常情况下要慢于音频处理的速度。文本驱动的方法可以看做音频驱动的扩展,在比较时与音频驱动有着相同的优劣势。

3 虚拟数字人的用户交互

随着人工智能的快速发展,虚拟数字人广泛应用于各种行业,例如智能客服、在线教育,拟人化的数字人可以在人机交互中显著增强用户体验。可交互的数字人通常需要包含多种人工智能框架(陶建华等,2022),并且在实时性上有着较高的要求。要想使用户在人机交互中获得较好的体验,数字人的设计者需要在提高效率、减小反应时间和反应的质量中做出权衡(陶建华等,2023)。

Shen 等人(2021)展示了 ViDA-MAN (visual dialog with digital humans), 一个用于多模态交互的数字人类代理,可以实时以音频和视觉方式回应即时语音查询,如图 17 所示。与传统的基于文本或语音的系统相比,ViDA-MAN 提供了类似人类的交互体验(例如生动的声音、自然的面部表情和身体手势)。ViDA-MAN 无缝集成了多模态技术,包括自动语音识别(automatic speech recognition, ASR)、多轮对话、文本转语音合成(TTS)、说话头部视频生成。凭借庞大的知识库支持,ViDA-MAN 能够与用户进行多个主题的聊天。

Zhen 等人(2022)根据数字人合成的驱动方法,将其分为化身类型、语音或文本内容驱动类型和人机交互类型。其中,人机交互类型结合了多种人工



图 17 多模态交互的数字人类代理 ViDA-MAN
(Shen 等, 2021)

Fig. 17 ViDA-MAN, a digital human agent for multimodal interaction (Shen et al., 2021)

智能技术,包括语音合成、识别、语义理解、多轮对话和知识图谱等。人机交互数字人可以快速识别用户真实意图,做出决策判断,自动匹配最佳答案,准确回答用户的问题,并实现流畅的人机对话。

Zhen 等人(2023)设计了人机交互系统框架,提出了一个系统的多模态人机交互框架,为说话头部生成模型的应用提供了新的思路,如图 18 所示。他们设计的系统主要包括语音模块、对话系统模块和说话头部生成。语音模块的自动语音识别(ASR)和文本转语音合成(TTS)分别对应于人类的听觉和语言功能;对话系统模块需要具备多轮对话的能力,满足用户进行闲聊的需求,用户的语音经过 ASR 后,问题传递给对话模块,对话模块必须根据用户的问题从知识库中检索或生成匹配的答案;说话头部生成模块的面部外观数据主要来自真实人物的照片、视频或混合形状(blendshape)角色模型系数。其中,语音模块和对话模块已经广泛商业化应用,可以满足人机交互的实时要求。但目前说话头部生成模型渲染和输出多模态视频仍需要较长时间。

由此可见,目前的可交互数字人通常由语音模块、对话系统模块和说话人生成模块构成。人机交互系统结构如图 19 所示。

3.1 数字人的语音模块

数字人的语音模块包括自动语音识别(ASR)和文本转语音合成(TTS)两部分。ASR 识别出人们的语音内容,该内容交由对话系统模块进行分析,对话系统模块作出相应的应答选择,应答文本再由 TTS 转换成语音输出。

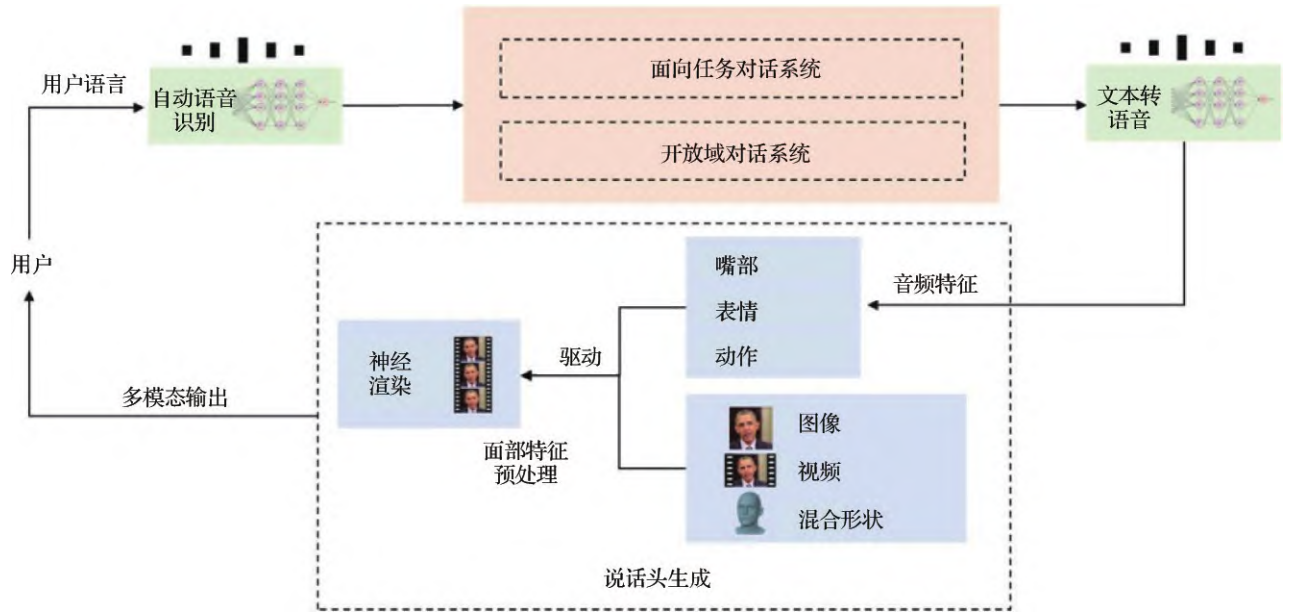


图 18 多模态人机交互框架(Zhen 等,2023)

Fig. 18 Multimodal human-computer interaction framework (Zhen et al. , 2023)

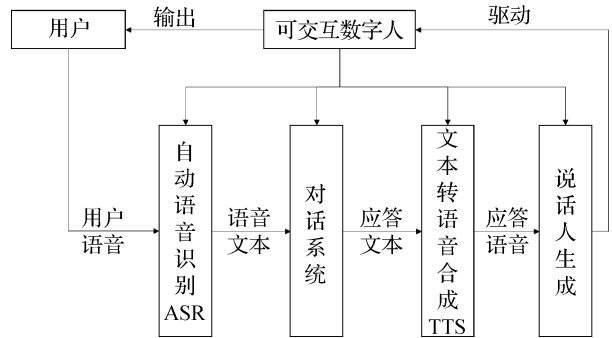


图 19 人机交互系统结构

Fig. 19 The system architecture of human-computer interaction

3. 1. 1 自动语音识别

Bengio 等人(2000)使用最基础的前馈神经网络(feedforward neural network, FNN)初步实现 ASR, 网络结构简单直观,能处理的文本序列长度受限于网络的输入长度。而后,Mikolov 等人(2010)基于循环神经网络(RNN)实现 ASR, RNN 与 FNN 相比,具有循环连接、使其在建模序列方面具有强大的能力, Amodei 等人(2015)通过使用神经网络替代传统的语音识别系统中的多个手工设计组件,能够处理包括嘈杂环境、口音和不同语言在内的多样化语音。但是传统 RNN 存在梯度消失和梯度爆炸问题。Sak 等人(2014)利用基于长短期记忆(long short-term memory, LSTM)的 RNN 架构进行大词汇量语音识别任务。对比了 LSTM 模型与传统 RNN 和深度神经网络(deep neural network, DNN)模型,在不同数量和配

置的参数下进行了训练和比较,证明了 LSTM 模型能较快收敛,并在相对较小的模型规模下实现了最先进的语音识别性能。

循环网络的训练时间较前馈网络更长,所以有研究人员选择改进 DNN 架构,赋予其记忆能力,以应对语音的上下文依赖关系。Peddinti 等人(2015)提出了一种时间延迟神经网络(time delay neural network, TDNN)架构,该架构在训练时间上与标准前馈深度网络 DNN 相当,同时能够建模长期时间依赖性。Povey 等人(2018)引入了一种 TDNN 的分解形式,称为 TDNN-F。TDNN-F 在结构上与经过 SVD(singular value decomposition)压缩的 TDNN 相同,但在训练过程中从随机起点开始,并要求每个矩阵的两个因子之一是半正交的。这样做可以显著改善 TDNN 的性能,并且与 TDNN- LSTM 混合模型的性能相当。

随着注意力机制(Vaswani 等,2017)的出现和发展,Transformer 也广泛运用于 ASR 任务。Zhang 等人(2020)使用 Transformer 代替 RNN 和 LSTM,由于 Transformer 结构相对简单,极大缩短了训练所需时间。Park 等人(2019)通过对输入进行数据增强,显著提高了 ASR 网络的性能。卷积神经网络(convolutional neural network, CNN)擅长利用局部特征,而 Transformer 更擅长捕获基于内容的全局交互, Conformer(Gulati 等,2020)将 CNN 和 Transformer 结

合起来,CNN作为Transformer的编码层,提取局部特征,期望模型兼具二者的优点。

近年来大模型火热,在ASR领域也有效果很好的大模型:OpenAI提出的语音识别系统Whisper(Radford等,2022)证明了扩大弱监督学习的规模,就可以得到强有力的语音识别模型。研究人员还发现对于足够大的模型,联合多语言和多任务训练没有缺点,甚至具有益处。

表3为自动语音识别模块总结。

表3 自动语音识别模块总结

Table 3 Summary of automatic speech recognition modules

主要结构	方法	特点
DNN	Bengio等人(2000)	文本长度受限
	Peddinti等人(2015)	DNN+记忆能力
	Povey等人(2018)	改善DNN性能
RNN+LSTM	Mikolov等人(2010)	纯RNN
	Sak等人(2014)	解决梯度消失和爆炸
	Amodei等人(2015)	CTC+RNN
Transformer	Zhang等人(2020)	显著缩短训练时间
	Park等人(2019)	数据增强
	Gulati等人(2020)	CNN+Transformer
混合模型	Shen等人(2021)	多模态实时交互
	Zhen等人(2023)	多模态实时交互
	Zhang等人(2022a)	易于使用

3.1.2 文本转语音合成

ASR广泛使用的RNN,LSTM模型同样也适用于TTS任务,Fan等人(2014)采用具有双向长短期记忆(bidirectional long short-term memory, BLSTM)单元的循环神经网络(RNN)来捕捉语音中任意两个时刻之间的相关或共现信息,用于参数化TTS合成。Mehri等人(2017)基于逐个音频样本生成的方法,该模型结合了无记忆模块(自回归多层感知器)和RNN,提出了SampleRNN,提升了模型性能。这些方法也在处理长序列时可能会遇到梯度消失或爆炸的问题。

改进DNN架构的代表方法有WaveNet(van den Oord等,2016),一种用于生成原始音频波形的深度神经网络。该模型是完全概率和自回归的,每个音

频样本的预测分布都依赖于之前的所有样本。WaveNet在应用于TTS时展现出先进的性能,其本身以及之后基于WaveNet的工作虽然能够合成高质量语音,但通常需要强大的GPU(graphics processing unit)来实现实时操作。van den Oord等人(2017)之后又提出了并行WaveNet,从一个训练好的WaveNet模型训练一个并行前馈网络,与传统WaveNet在感知质量上没有差异,并且相对于传统WaveNet,速度提升了1 000倍以上。Kim等人(2018)使用归一化流(normalizing flows)来建模原始音频数据,提出了FloWaveNet,其单阶段训练过程训练相比并行WaveNet更简单。Kalchbrenner等人(2018)将WaveNet简化为一个单层循环神经网络WaveRNN,减少了模型的复杂性和计算负担。

与ASR一致,基于Transformer的方法(Li等,2019)也极大程度提高了训练效率,同时解决了RNN难以建模长期依赖性的问题。Ren等人(2019)通过并行生成Mel频谱图,显著加快了语音合成的速度。Zeng等人(2020)通过一种新颖的对齐损失函数来学习文本和Mel频谱之间的对齐。基于生成流(generative flow)的方法(Miao等,2020;Kim等,2020)也被提出用于并行TTS,该方法构建非自回归TTS模型,相比于自回归模型实现了数量级的加速。

Valin和Skoglund(2019)提出的LPCNet(linear predictive coding net)则另辟蹊径,将数字信号处理(digital signal processing, DSP)和神经网络巧妙结合,能够显著提高语音合成的效率,使其能够在手机上部署。Valin等人(2022)对LPCNet进行改进,通过分层概率分布采样的算法改进,并结合计算改进,以更好地适应现有的CPU(central processing unit)架构,使得LPCNet算法可以在大多数现有手机上进行实时语音合成,并且合成质量也获得了提升。

Song等人(2021)提出了一种新颖的端到端语音合成方法,称为DIAN(duration informed auto-regressive network),由一个声学模型和一个独立的时长模型组成。其团队使用提出的DIAN方法为多说话人多风格语音克隆挑战(multi-speaker multi-style voice cloning challenge, M2VoC)开发了TTS系统,通过使用多说话人的时长模型和基于说话人的语音模型,该TTS系统能够克隆和生成每个新说话人或说话风格的语音。

TTS有时需要根据人们需求合成不同语调的语

音,目前,有两种常见的控制语音风格的方法:1)预定义一组语音风格,并使用分类指标表示不同语音风格。然而,这些模型在表现多样性方面存在限制,因为它们只能生成预定义的风格。2)使用参考语音作为风格输入,这导致一个问题,即提取的风格信息不直观或不可解释。Yang 等人(2023a)尝试使用自然语言作为风格提示来控制合成语音中的风格,构建了一个语音语料库,其中的语音样本不仅带有内容转录,还带有自然语言的风格描述,并提出了一种表达性TTS模型。音频风格控制示意图如图20所示。

来自百度的开源一体化语音工具包 PaddleSpeech(Zhang 等,2022a),旨在通过提供易于使用的命令行界面和简单的代码结构,促进语音处理技术的开发和研究。PaddleSpeech 可以完成 ASR 和

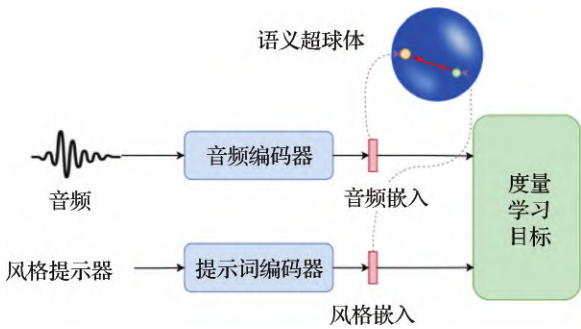


图 20 音频风格控制
Fig. 20 Audio style control

TTS 两种任务,大大降低了模型部署的复杂性,有利于人机交互数字人的构建。

表 4 为文本转语音合成模块总结。图 21 为自动语音识别模块/文本转语音合成模块方法的时间线。

表 4 文本转语音合成模块总结
Table 4 Summary of text-to-speech modules

主要结构	方法	特点
DNN	van den Oord 等人(2016)	WaveNet,高算力需求
	van den Oord 等人(2017)	并行 WaveNet,显著加速
	Kim 等人(2018)	Flow+WaveNet,训练更简单
RNN+LSTM	Fan 等人(2014)	双向长短期记忆单元 BLSTM
	Mehri 等人(2017)	MPL+RNN
	Kalchbrenner 等人(2018)	高度简化 WaveNet
	Valin 和 Skoglund(2019)	LPCNet,DSP+DNN
	Valin 等人(2022)	LPCNet,可在手机部署
Transformer	Li 等人(2019)	显著缩短训练时间
	Ren 等人(2019)	并行生成 Mel 频谱图
	Zeng 等人(2020)	并行生成 Mel 频谱图,采用对齐损失函数
混合模型	Miao 等人(2020); Kim 等人(2020)	Generative Flow+非自回归
	Shen 等人(2021)	多模态实时交互
	Zhen 等人(2023)	多模态实时交互
	Zhang 等人(2022a)	易于使用
	Song 等人(2021)	克隆生成新说话风格
	Yang 等人(2023a)	使用自然语言控制风格

3.2 数字人的对话系统模块

对话系统模块使得数字人能够对用户的话语进行应答、聊天和问答等人机交互。在用户的语音经过 ASR 后,问题被传递给对话模块。对话模块通常使用自然语言理解(natural language understanding,

NLU)单元,以识别用户的意图和相关实体信息。对话模块必须根据用户的问题从知识库中检索或生成匹配的答案。然而,在特定领域的多轮对话中,完全依靠模型生成答案是不可能的。在某些情况下,为了更好地考虑上下文信息,上述信息将被格式化为

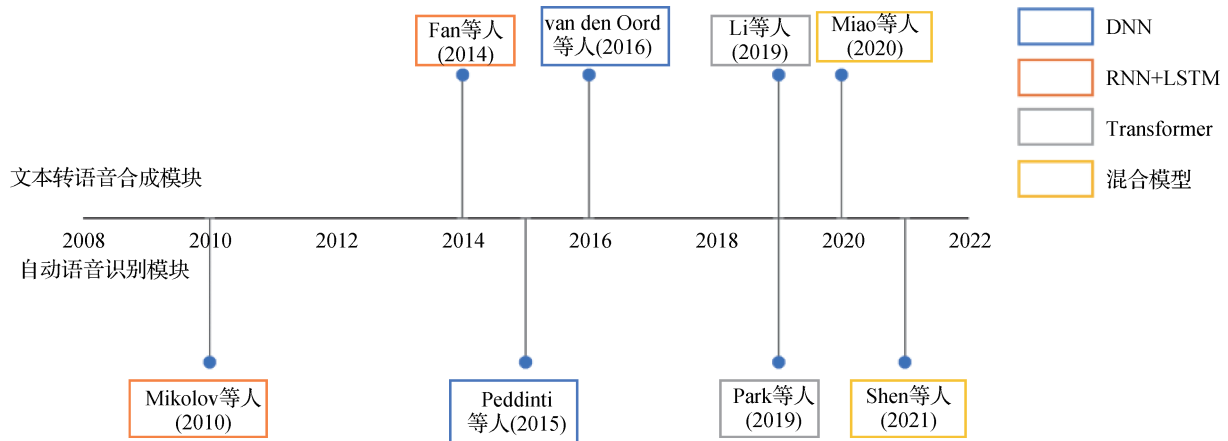


图21 自动语音识别模块/文本转语音合成模块方法时间线

Fig. 21 Timeline of automatic speech recognition/text-to-speech modules

对话状态跟踪器(dialog state tracking, DST)的输入, DST用于维护对话的当前状态。最后自然语言生成器(natural language generation, NLG)生成自然语言回复。例如 Kulkarni 等人(2016)提出分层强化学习(hierarchical deep Q-networks, h-DQN)框架, 顶层价值函数学习一个关于内在目标的策略, 而较低层的函数学习满足给定的目标所需要的原子动作策略, h-DQN 可用于选择下一个对话动作。

Ouyang 等人(2022)提出通过使用人类反馈进行微调的方法, 以在广泛的任务上使语言模型与用户意图保持一致。参照此方法训练出了极其强大的 ChatGPT(chat generative pre-trained transformer), 使用 ChatGPT 可以满足数字人的对话系统模块的绝大多数需求。

最后, 要得到与人交互的数字人, 还需要说话人生成模块的多模态驱动, 这一部分内容前文已经阐述。可见可交互的数字人包含了多种人工智能模型, 设计者需要设计合理的人机交互框架, 使得各个模块正常运作的同时又能正常交互。这也导致可交互数字人的实时性要求落在了每一个模块上, 要提高实时性, 需要提高每一个模块的效率。通过 Whisper、ChatGPT 等工作的成功, 可以认识到扩大训练规模具有的巨大潜力。

4 结 语

端到端的风格化数字人生成、驱动和交互系统能够打破传统的三维建模、动作采集、交互设计等环节之间的壁垒, 降低每个环节的操作难度, 为面向未

来的数字人应用提供傻瓜式解决方案。

风格化数字人的生成目前已可基于给定的照片或者视频, 结合深度学习和生成模型完成二维及三维高质量风格化人物的合成。虽然该类算法已广泛应用于社交媒体和影视创作, 但其仍然存在合成质量不稳定、风格化结果与预期不符等问题。因此, 如何提升现有算法, 实现高真实度、高可靠性的风格化, 仍有待继续深入研究。

自动驱动数字人进行自然的对话交流仍然是一项具有挑战性的任务。现有算法通过关键点、三维参数化模型、光流场、编码—解码结构等方法可以实现音频/文本/视频驱动的数字人合成。然而, 目前针对动态人脸驱动, 仍然存在无法合成高质量的嘴型、牙齿细节, 以及无法完成动态头发合成等问题。复杂的形貌和动态变化的几何形状, 给数字人驱动带来了难题。因此, 如何针对这些问题提出通用的解决方案, 还有待继续深入研究。

针对数字人的交互, 现有算法基于大语言模型可实现自然的数字人对话式交互, 但针对数字人与用户的多模态交互, 如动作、表情等的感知与反馈的相关研究还较少。如何将现有的对话数字人的相关研究拓展至可多维度交互的高真实度虚拟人还需要深入研究。

参考文献(References)

- Abdal R, Lee H Y, Zhu P H, Chai M L, Siarohin A, Wonka P and Tulyakov S. 2023. 3DAvatarGAN: bridging domains for personal-

- ized editable avatars//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 4552-4562 [DOI: 10.1109/CVPR52729.2023.00442]
- Almahairi A, Rajeshwar S, Sordani A, Bachman P and Courville A. 2018. Augmented CycleGAN: learning many-to-many mappings from unpaired data [EB/OL]. [2023-09-12]. <https://arxiv.org/pdf/1802.10151.pdf>
- Amodei D, Anubhai R, Battenberg E, Case C, Casper J, Catanzaro B, Chen J D, Chrzanowski M, Coates A, Diamos G, Elsen E, Engel J, Fan L X, Fougner C, Han T, Hannun A, Jun B, Legresley P, Lin L, Narang S, Ng A, Ozair S, Prenger R, Raiman J, Satheesh S, Seetapun D, Sengupta S, Wang Y, Wang Z Q, Wang C, Xiao B, Yogatama D, Zhan J and Zhu Z Y. 2015. Deep speech 2: end-to-end speech recognition in english and mandarin [EB/OL]. [2023-09-12]. <https://arxiv.org/pdf/1512.02595.pdf>
- Aneja S, Thies J, Dai A and Niessner M. 2023. ClipFace: text-guided editing of textured 3D morphable models//Proceedings of 2023 ACM SIGGRAPH Conference Proceedings. Los Angeles, USA: Association for Computing Machinery: #70 [DOI: 10.1145/3588432.3591566]
- Averbuch-Elor H, Cohen-Or D, Kopf J and Cohen M F. 2017. Bringing portraits to life. ACM Transactions on Graphics (TOG), 36(6): #196 [DOI: 10.1145/3130800.3130818]
- Baevski A, Zhou H, Mohamed A and Auli M. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations [EB/OL]. [2023-09-12]. <https://arxiv.org/pdf/2006.11477.pdf>
- Bengio Y, Ducharme R and Vincent P. 2000. A neural probabilistic language model//Proceedings of the 13th International Conference on Neural Information Processing Systems. Denver, USA: MIT Press: 893-899
- 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. [s.l.]: ACM Press/Addison-Wesley Publishing Co.: 187-194 [DOI: 10.1145/311535.311556]
- Blanz V and Vetter T. 2023. A morphable model for the synthesis of 3D faces//Seminal Graphics Papers: Pushing the Boundaries, Volume 2, 157-164
- 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. Vancouver, Canada: IEEE: 18392-18402 [DOI: 10.1109/CVPR52729.2023.01764]
- Cai Q, Ma M X, Wang C and Li H S. 2023. Image neural style transfer: a review. Computers and Electrical Engineering, 108: #108723 [DOI: 10.1016/j.compeleceng.2023.108723]
- Cao C, Simon T, Kim J K, Schwartz G, Zollhoefer M, Saito S S, Lombardi S, Wei S E, Belko D, Yu S I, Sheikh Y and Saragih J. 2022. Authentic volumetric avatars from a phone scan. ACM Transactions on Graphics (TOG), 41(4): #163 [DOI: 10.1145/3528223.3530143]
- Cao K D, Liao J and Yuan L. 2018. CariGANs: unpaired photo-to-caricature translation. ACM Transactions on Graph ICS (TOG), 37(6): #244 [DOI: 10.1145/3272127.3275046]
- Cao Y, Tien W C, Faloutsos P and Pighin F. 2005. Expressive speech-driven facial animation. ACM Transactions on Graphics (TOG), 24(4): 1283-1302 [DOI: 10.1145/1095878.1095881]
- 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. New Orleans, USA: IEEE: 16102-16112 [DOI: 10.1109/CVPR52688.2022.01565]
- Chatziagapi A, Athar S, Jain A, Rohith M V, Bhat V and Samaras D. 2023. LipNeRF: what is the right feature space to lip-sync a NeRF?//Proceedings of the 17th IEEE International Conference on Automatic Face and Gesture Recognition (FG). Waikoloa Beach, USA: IEEE: 1-8 [DOI: 10.1109/FG57933.2023.10042567]
- Chen D D, Liao J, Yuan L, Yu N H and Hua G. 2017. Coherent online video style transfer//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 1114-1123 [DOI: 10.1109/Iccv.2017.126]
- Chen L L, Li Z H, Maddox R K, Duan Z Y and Xu C L. 2018. Lip movements generation at a glance//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 538-553 [DOI: 10.1007/978-3-030-01234-2_32]
- 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. Long Beach, USA: IEEE: 7824-7833 [DOI: 10.1109/CVPR.2019.00802]
- Chen Z, Xu X D, Yan Y C, Pan Y, Zhu W H, Wu W, Dai B and Yang X K. 2023. HyperStyle3D: text-guided 3D portrait stylization via hypernetworks [EB/OL]. [2023-09-12]. <https://arxiv.org/pdf/2304.09463.pdf>
- 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. Long Beach, USA: IEEE: 10093-10103 [DOI: 10.1109/CVPR.2019.01034]
- Das D, Biswas S, Sinha S and Bhowmick B. 2020. Speech-driven facial animation using cascaded gans for learning of motion and texture//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 408-424 [DOI: 10.1007/978-3-030-58577-8_25]
- Edwards P, Landreth C, Fiume E and Singh K. 2016. JALI: an animator-centric viseme model for expressive lip synchronization. ACM Transactions on Graphics (TOG), 35(4): #127 [DOI: 10.1145/2897824.2925984]

- Eskimez S E, Zhang Y and Duan Z Y. 2022. Speech driven talking face generation from a single image and an emotion condition. *IEEE Transactions on Multimedia*, 24: 3480-3490 [DOI: 10.1109/TMM.2021.3099900]
- Ezzat T and Poggio T. 2002. MikeTalk: a talking facial display based on morphing visemes//*Proceedings Computer Animation '98*. Philadelphia, USA: IEEE: 96-102 [DOI: 10.1109/CA.1998.681913]
- Fan Y C, Qian Y, Xie F L and Soong F K. 2014. TTS synthesis with bidirectional LSTM based recurrent neural networks//*Interspeech 2014*. Singapore, Singapore: ISCA: 1964-1968 [DOI: 10.21437/interspeech.2014-443]
- Fan Y R, Lin Z J, Saito J, Wang W P and Komura T. 2022. FaceFormer: speech-driven 3D facial animation with transformers//*Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, USA: IEEE: 18749-18758 [DOI: 10.1109/CVPR52688.2022.01821]
- Fried O, Tewari A, Zollhöfer M, Finkelstein A, Shechtman E, Goldman D B, Genova K, Jin Z Y, Theobalt C and Agrawala M. 2019. Text-based editing of talking-head video. *ACM Transactions on Graphics (TOG)*, 38(4): #68 [DOI: 10.1145/3306346.3323028]
- Gal R, Patashnik O, Maron H, Chechik G and Cohen-Or D. 2021. StyleGAN-NADA: CLIP-guided domain adaptation of image generators [EB/OL]. [2023-09-12]. <https://arxiv.org/pdf/2108.00946.pdf>
- Gao W, Lie Y J, Yin Y H and Yang M H. 2020. Fast video multi-style transfer//*Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Snowmass, USA: IEEE: 3211-3219 [DOI: 10.1109/WACV45572.2020.9093420]
- Gatys L A, Ecker A S and Bethge M. 2016. Image style transfer using convolutional neural networks//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, USA: IEEE: 2414-2423 [DOI: 10.1109/Cvpr.2016.265]
- Geng J H, Shao T J, Zheng Y Y, Weng Y L and Zhou K. 2018. Warp-guided gans for single-photo facial animation. *ACM Transactions on Graphics (TOG)*, 37(6): #231 [DOI: 10.1145/3272127.3275043]
- Genova K, Cole F, Maschinot A, Sarna A, Vlastic D and Freeman W T. 2018. Unsupervised training for 3D morphable model regression//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, USA: IEEE: 8377-8386 [DOI: 10.1109/CVPR.2018.00874]
- Gulati A, Qin J, Chiu C C, Parmar N, Zhang Y, Yu J H, Han W, Wang S B, Zhang Z D, Wu Y H and Pang R M. 2020. Conformer: convolution-augmented transformer for speech recognition [EB/OL]. [2023-09-12]. <https://arxiv.org/pdf/2005.08100.pdf>
- Guo Y D, Chen K Y, Liang S, Liu Y J, Bao H J and Zhang J Y. 2021. AD-NeRF: audio driven neural radiance fields for talking head synthesis//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*. Montreal, Canada: IEEE: 5764-5774 [DOI: 10.1109/ICCV48922.2021.00573]
- Guo Y D, Jiang L, Cai L and Zhang J Y. 2019. 3D magic mirror: automatic video to 3D caricature translation [EB/OL]. [2023-09-12]. <https://arxiv.org/pdf/1906.00544.pdf>
- Ha S, Kersner M, Kim B, Seo S and Kim D. 2020. Marionette: few-shot face reenactment preserving identity of unseen targets//*Proceedings of the 34th AAAI Conference on Artificial Intelligence*. New York, USA: AAAI: 10893-10900 [DOI: 10.1609/AAAI.V34I07.6721]
- Han F Z, Ye S Q, He M M, Chai M L and Liao J. 2023. Exemplar-based 3D portrait stylization. *IEEE Transactions on Visualization and Computer Graphics*, 29(2): 1371-1383 [DOI: 10.1109/TVCG.2021.3114308]
- Hao C H, Du Y Y, Wang L and Wang B B. 2024. Survey of digital face rendering and appearance recovery methods. *Journal of Image and Graphics*, 29(9): 2513-2540 (郝琮晖, 杜悠扬, 王璐, 王贝贝). 2024. 数字人脸渲染与外观恢复方法综述. *中国图象图形学报*, 29(9): 2513-2540 [DOI: 10.11834/jig.230683]
- Haque A, Tancik M, Efros A A, Holynski A and Kanazawa A. 2023. Instruct-NeRF2NeRF: editing 3D scenes with instructions [EB/OL]. [2023-09-12]. <https://arxiv.org/pdf/2303.12789.pdf>
- Ho J, Salimans T, Gritsenko A, Chan W, Norouzi M and Fleet D J. 2022. Video diffusion models [EB/OL]. [2023-09-12]. <https://arxiv.org/pdf/2204.03458.pdf>
- Hu L, Qi J W, Zhang B, Pan P and Xu Y H. 2021. Text-driven 3D avatar animation with emotional and expressive behaviors//*Proceedings of the 29th ACM International Conference on Multimedia*. Virtual Event, China: ACM: 2816-2818 [DOI: 10.1145/3474085.3478569]
- Jamaludin A, Chung J S and Zisserman A. 2019. You said that?: synthesizing 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 Q Y, Wu W, Xu F and Cao X. 2022. EAMM: one-shot emotional talking face via audio-based emotion-aware motion model//*Proceedings of 2022 ACM SIGGRAPH Conference Proceedings*. Vancouver, Canada: ACM: #61 [DOI: 10.1145/3528233.3530745]
- 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*. Nashville, USA: IEEE: 14075-14084 [DOI: 10.1109/CVPR46437.2021.01386]
- Jiang K W, Chen S Y, Liu F L, Fu H B and Gao L. 2022. NeRFFaceEditing: disentangled face editing in neural radiance fields//*Proceedings of 2022 SIGGRAPH Asia Conference Papers*. Daegu, Republic of Korea: Association for Computing Machinery: #31 [DOI: 10.1145/3550469.3555377]
- Kalberer G A and Van Gool L. 2001. Face animation based on observed 3D speech dynamics//*Proceedings of the 14th Conference on Computer Animation*. Seoul, Korea (South): IEEE: 20-27 [DOI: 10.1109/CA.2001.982373]

- Kalchbrenner N, Elsen E, Simonyan K, Noury S, Casagrande N, Lockhart E, Stimberg F, van den Oord A, Dieleman S and Kavukcuoglu K. 2018. Efficient neural audio synthesis [EB/OL]. [2023-09-12]. <https://arxiv.org/pdf/1802.08435.pdf>
- Karras T, Aila T, Laine S, Herva A and Lehtinen J. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4): #94 [DOI: 10.1145/3072959.3073658]
- Karras T, Laine S and Aila T. 2019. A style-based generator architecture for generative adversarial networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA; IEEE: 4396-4405 [DOI: 10.1109/CVPR.2019.00453]
- Kim H, Elgharib M, Zollhöfer M, Seidel H P, Beeler T, Richardt C and Theobalt C. 2019. Neural style-preserving visual dubbing. *ACM Transactions on Graphics (TOG)*, 38(6): #178 [DOI: 10.1145/3355089.3356500]
- Kim J, Kim S, Kong J and Yoon S. 2020. Glow-TTS: a generative flow for text-to-speech via monotonic alignment search//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada; Curran Associates Inc.: 8067-8077
- Kim S, Lee S G, Song J, Kim J and Yoon S. 2018. FloWaveNet: a generative flow for raw audio [EB/OL]. [2023-09-12]. <https://arxiv.org/pdf/1811.02155.pdf>
- Kulkarni T D, Narasimhan K R, Saeedi A and Tenenbaum J B. 2016. Hierarchical deep reinforcement learning: integrating temporal abstraction and intrinsic motivation [EB/OL]. [2024-09-07]. <https://arxiv.org/pdf/1604.06057.pdf>
- Kumar R, Sotelo J, Kumar K, De Brébisson A and Bengio Y. 2017. ObamaNet: photo-realistic lip-sync from text [EB/OL]. [2024-09-07]. <https://arxiv.org/pdf/1801.01442.pdf>
- Li J H, Zhang J W, Bai X, Zhou J and Gu L. 2023. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France; IEEE: 7534-7544 [DOI: 10.1109/ICCV51070.2023.00696]
- Li L C, Wang S Z, Zhang Z M, Ding Y, Zheng Y X, Yu X and Fan C J. 2021. Write-a-speaker: text-based emotional and rhythmic talking-head generation//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtually; AAAI: 1911-1920 [DOI: 10.1609/AAAI.V35I3.16286]
- Li N H, Liu S J, Liu Y Q, Zhao S and Liu M. 2019. Neural speech synthesis with transformer network//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA; AAAI: 6706-6713 [DOI: 10.1609/aaai.v33i01.33016706]
- Li S X. 2023. Instruct-Video2Avatar: video-to-avatar generation with instructions [EB/OL]. [2024-09-07]. <https://arxiv.org/pdf/2306.02903.pdf>
- Li S X and Pan Y. 2023. Rendering and reconstruction based 3D portrait stylization//Proceedings of 2023 IEEE International Conference on Multimedia and Expo (ICME). Brisbane, Australia; IEEE: 912-917 [DOI: 10.1109/ICME55011.2023.00161]
- 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]
- Lin J K, Yuan Y and Zou Z X. 2021. MeInGame: create a game character face from a single portrait//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtual Event; AAAI: 311-319 [DOI: 10.1609/AAAI.v35i1.16106]
- Liu A A, Su Y T, Wang L J, Li B, Qian Z X, Zhang W M, Zhou L N, Zhang X P, Zhang Y D, Huang J W and Yu N H. 2024. Review on the progress of the AIGC visual content generation and traceability. *Journal of Image and Graphics*, 29(6): 1535-1554 (刘安安, 苏育挺, 王岚君, 李斌, 钱振兴, 张卫明, 周琳娜, 张新鹏, 张勇东, 黄继武, 俞能海. 2024. AIGC 视觉内容生成与溯源研究进展. 中国图象图形学报, 29(6): 1535-1554) [DOI: 10.11834/jig.240003]
- Liu J L, Zhu Z Y, Ren Y, Huang W C, Huai B X, Yuan N and Zhao Z. 2022a. Parallel and high-fidelity text-to-lip generation//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtual Event; AAAI: 1738-1746 [DOI: 10.1609/AAAI.V36I2.20066]
- Liu N, Li S, Du Y L, Torralba A and Tenenbaum J B. 2022b. Compositional visual generation with composable diffusion models//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel; Springer: 423-439 [DOI: 10.1007/978-3-031-19790-1_26]
- Liu X, Xu Y H, Wu Q Y, Zhou H, Wu W and Zhou B L. 2022c. 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]
- Mehri S, Kumar K, Gulrajani I, Kumar R, Jain S, Sotelo J, Courville A and Bengio Y. 2017. SampleRNN: an unconditional end-to-end neural audio generation model [EB/OL]. [2023-09-12]. <https://arxiv.org/pdf/1612.07837.pdf>
- Miao C F, Liang S, Chen M C, Ma J, Wang S J and Xiao J. 2020. Flow-TTS: a non-autoregressive network for text to speech based on flow//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain; IEEE: 7209-7213 [DOI: 10.1109/icassp40776.2020.9054484]
- Mikolov T, Karafiát M, Burget L, Černocký J and Khudanpur S. 2010. Recurrent neural network based language model//Interspeech 2010. Makuhari, Japan; ISCA: 1045-1048 [DOI: 10.21437/interspeech.2010-343]
- Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R and Ng R. 2021. NeRF: representing scenes as neural radiance

- fields for view synthesis. *Communications of the ACM*, 65(1): 99-106 [DOI: 10.1145/3503250]
- Müller T, Evans A, Schied C and Keller A. 2022. Instant neural graphics primitives with a multiresolution hash encoding [EB/OL]. [2023-09-12]. <https://arxiv.org/pdf/2201.05989.pdf>
- Nguyen-Phuoc T, Schwartz G, Ye Y T, Lombardi S and Xiao L. 2023. AlteredAvatar: stylizing dynamic 3D avatars with fast style adaptation [EB/OL]. [2024-09-07]. <https://arxiv.org/pdf/2305.19245v1.pdf>
- Olivier N, Kerbiriou G, Arguelaguet F, Avril Q, Danieau F, Guillotel P, Hoyet L and Multon F. 2022. Study on automatic 3D facial caricaturization: from rules to deep learning. *Frontiers in Virtual Reality*, 2: #785104 [DOI: 10.3389/FRVIR.2021.785104]
- Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C L, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kellton F, Miller L, Simens M, Askeel A, Welinder P, Christiano P, Leike J and Lowe R. 2022. Training language models to follow instructions with human feedback [EB/OL]. [2024-09-07]. <https://arxiv.org/pdf/2203.02155.pdf>
- Park D S, Chan W, Zhang Y, Chiu C C, Zoph B, Cubuk E D and Le Q V. 2019. SpecAugment: a simple data augmentation method for automatic speech recognition [EB/OL]. [2023-09-12]. <https://arxiv.org/pdf/1904.08779.pdf>
- Park S J, Kim M, Hong J, Choi J and Ro Y M. 2022. SyncTalkFace: talking face generation with precise lip-syncing via audio-lip memory//*Proceedings of the 36th AAAI Conference on Artificial Intelligence*. Virtual Event: AAAI; 2062-2070 [DOI: 10.1609/AAAI.V36I2.20102]
- Peddinti V, Povey D and Khudanpur S. 2015. A time delay neural network architecture for efficient modeling of long temporal contexts//*Interspeech 2015*. Dresden, Germany: ISCA; 3214-3218 [DOI: 10.21437/interspeech.2015-647]
- Povey D, Cheng G F, Wang Y M, Li K, Xu H N, Yarmohammadi M and Khudanpur S. 2018. Semi-orthogonal low-rank matrix factorization for deep neural networks//*Interspeech 2018*. Hyderabad, India: ISCA; 3743-3747 [DOI: 10.21437/interspeech.2018-1417]
- Prajwal K R, Mukhopadhyay R, Philip J, Jha A, Namboodiri V and Jawahar C. 2019. Towards automatic face-to-face translation//*Proceedings of the 27th ACM International Conference on Multimedia*. New York, USA: ACM; 1428-1436 [DOI: 10.1145/3343031.3351066]
- 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, Xu T, Brockman G, McLeavey C and Sutskever I. 2022. Robust speech recognition via large-scale weak supervision [EB/OL]. [2024-09-07]. <https://arxiv.org/pdf/2212.04356.pdf>
- Ren Y, Ruan Y J, Tan X, Qin T, Zhao S, Zhao Z and Liu T Y. 2019. FastSpeech: fast, robust and controllable text to speech [EB/OL]. [2023-09-12]. <https://arxiv.org/pdf/1905.09263.pdf>
- 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. 2022. High-resolution image synthesis with latent diffusion models//*Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, USA: IEEE; 10674-10685 [DOI: 10.1109/Cvpr52688.2022.01042]
- Ruder M, Dosovitskiy A and Brox T. 2016. Artistic style transfer for videos//*Proceedings of the 38th German Conference on Pattern Recognition*. Hannover, Germany: Springer; 26-36 [DOI: 10.1007/978-3-319-45886-1_3]
- Sadoughi N and Busso C. 2021. Speech-driven expressive talking lips with conditional sequential generative adversarial networks. *IEEE Transactions on Affective Computing*, 12(4): 1031-1044 [DOI: 10.1109/TAFFC.2019.2916031]
- Sak H, Senior A and Beaufays F. 2014. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition [EB/OL]. [2024-09-07]. <https://arxiv.org/pdf/1402.1128.pdf>
- Schneider S, Baevski A, Collobert R and Auli M. 2019. Wav2vec: unsupervised pre-training for speech recognition//*Interspeech 2019*. Graz, Austria: ISCA; 3465-3469 [DOI: 10.21437/Interspeech.2019-1873]
- Shao R Z, Sun J X, Peng C, Zheng Z R, Zhou B Y, Zhang H W and Liu Y B. 2023a. Control4D: efficient 4D portrait editing with text [EB/OL]. [2024-09-07]. <https://arxiv.org/pdf/2305.20082.pdf>
- Shao R Z, Zheng Z R, Tu H Z, Liu B N, Zhang H W and Liu Y B. 2023b. Tensor4D: efficient neural 4D decomposition for high-fidelity dynamic reconstruction and rendering//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver, Canada: IEEE; 16632-16642 [DOI: 10.1109/CVPR52729.2023.01596]
- 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 T, Zuo J W, Shi F, Zhang J, Jiang L Q, Chen M, Zhang Z C, Zhang W, He X D and Mei T. 2021. ViDA-MAN: visual dialog with digital humans//*Proceedings of the 29th ACM International Conference on Multimedia*. Virtual Event, China: ACM; 2789-2791 [DOI: 10.1145/3474085.3478560]
- Shi T Y, Yuan Y, Fan C J, Zou Z X, Shi Z W and Liu Y. 2019. Face-to-

- parameter translation for game character auto-creation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 161-170 [DOI: 10.1109/ICCV.2019.00025]
- Siarohin A, Lathuilière S, Tulyakov S, Ricci E and Sebe N. 2020. First order motion model for image animation [EB/OL]. [2023-09-12]. <https://arxiv.org/pdf/2003.00196.pdf>
- Sohl-Dickstein J, Weiss E A, Maheswaranathan N and Ganguli S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics//Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille, France: JMLR.org: 2256-2265
- Song H K, Woo S H, Lee J, Yang S, Cho H, Lee Y, Choi D and Kim K W. 2022a. Talking face generation with multilingual tts//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 21393-21398 [DOI: 10.1109/CVPR52688.2022.02074]
- Song L S, Wu W, Qian C, He R and Loy C C. 2022b. Everybody's talkin': let me talk as you want. IEEE Transactions on Information Forensics and Security, 17: 585-598 [DOI: 10.1109/TIFS.2022.3146783]
- Song W, Yuan X, Zhang Z C, Zhang C, Wu Y Z, He X D and Zhou B W. 2021. Dian: duration informed auto-regressive network for voice cloning//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada: IEEE: 8598-8602 [DOI: 10.1109/icassp39728.2021.9414727]
- Song Y, Zhu J W, Li D W, Wang X and Qi H R. 2019. Talking face generation by conditional recurrent adversarial network [EB/OL]. [2024-09-07]. <https://arxiv.org/pdf/1804.04786.pdf>
- Sun Y Q, He R, Tan W M and Yan B. 2023. Instruct-neuraltalker: editing audio-driven talking radiance fields with instructions [EB/OL]. [2024-09-07]. <https://arxiv.org/pdf/2306.10813.pdf>
- 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]. [2023-09-12]. <https://arxiv.org/pdf/2211.12368.pdf>
- Tao J H, Fan C H, Lian Z, Lyu Z, Shen Y and Liang S. 2024. Development of multimodal sentiment recognition and understanding. Journal of Image and Graphics, 29(6): 1607-1627 (陶建华, 范存航, 连政, 吕钊, 沈莹, 梁山. 2024. 多模态情感识别与理解发展现状及趋势. 中国图象图形学报, 29(6): 1607-1627) [DOI: 10.11834/jig.240017]
- Tao J H, Gong J T, Gao N, Fu S W, Liang S and Yu C. 2023. Human-computer interaction for virtual-real fusion. Journal of Image and Graphics, 28(6): 1513-1542 (陶建华, 龚江涛, 高楠, 傅四维, 梁山, 喻纯. 2023. 面向虚实融合的人机交互. 中国图象图形学报, 28(6): 1513-1542) [DOI: 10.11834/jig.230020]
- Tao J H, Wu Y C, Yu C, Weng D D, Li G J, Han T, Wang Y T and Liu B. 2022. A survey on multi-modal human-computer interaction. Journal of Image and Graphics, 27(6): 1956-1987 (陶建华, 巫英才, 喻纯, 翁冬冬, 李冠君, 韩腾, 王运涛, 刘斌. 2022. 多模态人机交互综述. 中国图象图形学报, 27(6): 1956-1987) [DOI: 10.11834/jig.220151]
- 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—ECCV 2020. Glasgow, UK: Springer: 716-731 [DOI: 10.1007/978-3-030-58517-4_42]
- Tian G Z, Yuan Y and Liu Y. 2019. Audio2face: generating speech/face animation from single audio with attention-based bidirectional LSTM networks//Proceedings of 2019 IEEE International Conference on Multimedia and Expo Workshops (ICMEW). Shanghai, China: IEEE: 366-371 [DOI: 10.1109/ICMEW.2019.00069]
- Valin J M, Isik U, Smaragdis P and Krishnaswamy A. 2022. Neural speech synthesis on a shoestring: improving the efficiency of LPC-Net//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE: 8437-8441 [DOI: 10.1109/icassp43922.2022.9746103]
- Valin J M and Skoglund J. 2019. LPCNET: improving neural speech synthesis through linear prediction//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE: 5891-5895 [DOI: 10.1109/icassp.2019.8682804]
- van den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A and Kavukcuoglu K. 2016. Wavenet: a generative model for raw audio [EB/OL]. [2024-09-07]. <https://arxiv.org/pdf/1609.03499.pdf>
- van den Oord A, Li Y Z, Babuschkin I, Simonyan K, Vinyals O, Kavukcuoglu K, van den Driessche G, Lockhart E, Cobo L C, Stimberg F, Casagrande N, Grewe D, Noury S, Dieleman S, Elsen E, Kalchbrenner N, Zen H, Graves A, King H, Walters T, Belov D and Hassabis D. 2017. Parallel WaveNet: fast high-fidelity speech synthesis [EB/OL]. [2023-09-12]. <https://arxiv.org/pdf/1711.10433.pdf>
- 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
- Verma A, Rajput N and Subramaniam L V. 2003. Using viseme based acoustic models for speech driven lip synthesis//Proceedings of 2003 International Conference on Multimedia and Expo. Baltimore, USA: IEEE: 533-536 [DOI: 10.1109/ICME.2003.1221366]
- Wang C, Jiang R X, Chai M L, He M M, Chen D D and Liao J. 2024. NeRF-Art: text-driven neural radiance fields stylization. IEEE Transactions on Visualization and Computer Graphics, 30(8): 4983-4996 [DOI: 10.1109/TVCG.2023.3283400]

- Wang H, Lin G S, Hoi S C H and Miao C Y. 2022a. 3D cartoon face generation with controllable expressions from a single GAN image [EB/OL]. [2024-09-07]. <https://arxiv.org/pdf/2207.14425.pdf>
- Wang K S Y, Wu Q Y, Song L S, Yang Z Q, Wu W, Qian C, He R, Qiao Y and Loy C C. 2020. MEAD: a large-scale audio-visual dataset for emotional talking-face generation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 700-717 [DOI: 10.1007/978-3-030-58589-1_42]
- Wang S Z, Li L C, Ding Y, Fan C J and Yu X. 2021. Audio2Head: audio-driven one-shot talking-head generation with natural head motion//Proceedings of the 30th International Joint Conference on Artificial Intelligence. Virtual Event: IJCAI: 1098-1105 [DOI: 10.24963/IJCAI.2021/152]
- Wang S Z, Li L C, Ding Y and Yu X. 2022b. One-shot talking face generation from single-speaker audio-visual correlation learning//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtually: AAAI: 2531-2539 [DOI: 10.1609/AAAI.V36I3.20154]
- Wang S Z, Zeng W H, Wang X, Yang H, Chen L, Zhang C, Wu M, Yuan Y, Zeng Y, Zheng M and Liu J. 2023. SwiftAvatar: efficient auto-creation of parameterized stylized character on arbitrary avatar engines//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI: 6101-6109 [DOI: 10.1609/AAAI.v37i5.25753]
- Wen X, Wang M, Richardt C, Chen Z Y and Hu S M. 2020. Photorealistic audio-driven video portraits. IEEE Transactions on Visualization and Computer Graphics, 26(12): 3457-3466 [DOI: 10.1109/TVCG.2020.3023573]
- Wiles O, Koepke A S and Zisserman A. 2018. X2Face: a network for controlling face generation using images, audio, and pose codes//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 690-706 [DOI: 10.1007/978-3-030-01261-8_41]
- Wood E, Baltrušaitis T, Hewitt C, Johnson M, Shen J J, Milosavljević N, Wilde D, Garbin S, Sharp T, Stojiljković I, Cashman T and Valentin J. 2022. 3D face reconstruction with dense landmarks//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 160-177 [DOI: 10.1007/978-3-031-19778-9_10]
- Wu J Z, Ge Y X, Wang X T, Lei W X, Gu Y C, Shi Y F, Hsu W, Shan Y, Qie X H and Shou M Z. 2023. Tune-A-Video: one-shot tuning of image diffusion models for text-to-video generation [EB/OL]. [2023-09-12]. <https://arxiv.org/pdf/2212.11565.pdf>
- Wu Z Z, Watts O and King S. 2016. Merlin: an open source neural network speech synthesis system//9th ISCA Workshop on Speech Synthesis Workshop. Sunnyvale, USA: [s.n.]: 202-207 [DOI: 10.21437/SSW.2016-33]
- Yang D C, Liu S X, Huang R J, Weng C and Meng H L. 2023a. InstructTTS: modelling expressive TTS in discrete latent space with natural language style prompt [EB/OL]. [2024-09-07]. <https://arxiv.org/pdf/2301.13662.pdf>
- Yang S, Zhou Y F, Liu Z W and Loy C C. 2023b. Rerender a video: zero-shot text-guided video-to-video translation//Proceedings of 2023 SIGGRAPH Asia Conference Papers. Sydney, Australia: ACM: #95 [DOI: 10.1145/3610548.3618160]
- Yao S Y, Zhong R Z, Yan Y C, Zhai G T and Yang X K. 2022. DFA-NeRF: personalized talking head generation via disentangled face attributes neural rendering [EB/OL]. [2023-09-12]. <https://arxiv.org/pdf/2201.00791.pdf>
- Ye Z H, Jiang Z Y, Ren Y, Liu J L, He J Z and Zhao Z. 2023a. GeneFace: generalized and high-fidelity audio-driven 3D talking face synthesis [EB/OL]. [2023-09-12]. <https://arxiv.org/pdf/2301.13430.pdf>
- Ye Z P, Xia M F, Sun Y N, Yi R, Yu M J, Zhang J Y, Lai Y K and Liu Y J. 2023b. 3D-CariGAN: an end-to-end solution to 3D caricature generation from normal face photos. IEEE Transactions on Visualization and Computer Graphics, 29(4): 2203-2210 [DOI: 10.1109/tvcg.2021.3126659]
- Yi R, Ye Z P, Zhang J Y, Bao H J and Liu Y J. 2020. Audio-driven talking face video generation with learning-based personalized head pose [EB/OL]. [2024-09-07]. <https://arxiv.org/pdf/2002.10137.pdf>
- Yin F, Zhang Y, Cun X, Cao M D, Fan Y B, Wang X, Bai Q Y, Wu B Y, Wang J and Yang Y J. 2022. StyleHEAT: one-shot high-resolution editable talking face generation via pre-trained StyleGAN//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 85-101 [DOI: 10.1007/978-3-031-19790-1_6]
- Yu L Y, Yu J and Ling Q. 2019. Mining audio, text and visual information for talking face generation//Proceedings of 2019 IEEE International Conference on Data Mining (ICDM). Beijing, China: IEEE: 787-795 [DOI: 10.1109/ICDM.2019.00089]
- Zaremba W, Sutskever I and Vinyals O. 2015. Recurrent neural network regularization [EB/OL]. [2024-09-07]. <https://arxiv.org/pdf/1409.2329.pdf>
- Zeng Z, Wang J Z, Cheng N, Xia T and Xiao J. 2020. Aligntts: efficient feed-forward text-to-speech system without explicit alignment//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE: 6714-6718 [DOI: 10.1109/icassp40776.2020.9054119]
- Zhang B W, Qi C Y, Zhang P, Zhang B, Wu H, Chen D, Chen Q F, Wang Y and Wen F. 2023a. MetaPortrait: identity-preserving talking head generation with fast personalized adaptation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE: 22096-22105 [DOI: 10.1109/CVPR52729.2023.02116]
- Zhang C X, Zhao Y F, Huang Y F, Zeng M, Ni S F, Budagavi M and Guo X H. 2021. FACIAL: synthesizing dynamic talking face with implicit attribute learning//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE:

- 3847-3856 [DOI: 10.1109/ICCV48922.2021.00384]
- Zhang H, Yuan T, Chen J K, Li X T, Zheng R J, Huang Y X, Chen X J, Gong E L, Chen Z Y, Hu X G, Yu D H, Ma Y J and Huang L. 2022a. PaddleSpeech: an easy-to-use all-in-one speech toolkit [EB/OL]. [2024-09-07]. <https://arxiv.org/pdf/2205.12007.pdf>
- Zhang L M, Rao A Y and Agrawala M. 2023b. Adding conditional control to text-to-image diffusion models [EB/OL]. [2024-09-07]. <https://arxiv.org/pdf/2302.05543.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. 2023c. DreamFace: progressive generation of animatable 3D faces under text guidance [EB/OL]. [2023-09-12]. <https://arxiv.org/pdf/2304.03117.pdf>
- Zhang Q, Lu H, Sak H, Tripathi A, McDermott E, Koo S and Kumar S. 2020. Transformer transducer: a streamable speech recognition model with transformer encoders and RNN-T loss//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE: 7829-7833 [DOI: 10.1109/icassp40776.2020.9053896]
- Zhang W X, Cun X, Wang X, Zhang Y, Shen X, Guo Y, Shan Y and Wang F. 2023d. SadTalker: learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 8652-8661 [DOI: 10.1109/CVPR52729.2023.00836]
- Zhang Y H, He W H, Li M L, Tian K, Zhang Z Y, Cheng J, Wang Y Y and Liao J X. 2022c. Meta talk: learning to data-efficiently generate audio-driven lip-synchronized talking face with high definition//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE: 4848-4852 [DOI: 10.1109/ICASSP43922.2022.9747284]
- Zhen R, Song W C and Cao J. 2022. Research on the application of virtual human synthesis technology in human-computer interaction//Proceedings of the 22nd IEEE/ACIS International Conference on Computer and Information Science (ICIS). Zhuhai, China: IEEE: 199-204 [DOI: 10.1109/ICIS54925.2022.9882355]
- Zhen R, Song W C, He Q, Cao J, Shi L and Luo J. 2023. Human-computer interaction system: a survey of talking-head generation. Electronics, 12(1): #218 [DOI: 10.3390/electronics12010218]
- Zhou H, Liu Y, Liu Z W, Luo P and Wang X G. 2019. Talking face generation by adversarially disentangled audio-visual representation//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI: 9299-9306 [DOI: 10.1609/AAAI.V33I01.33019299]
- Zhou H, Sun Y S, Wu W, Loy C C, 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. 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. MakeltTalk: speaker-aware talking-head animation. ACM Transactions on Graphics (TOG), 39(6): #221 [DOI: 10.1145/3414685.3417774]
- Zielonka W, Bolkart T and Thies J. 2023. Instant volumetric head avatars//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 4574-4584 [DOI: 10.1109/CVPR52729.2023.00444]

作者简介

潘焱,女,副教授,主要研究方向为虚拟现实、人机交互和角色动画。E-mail: whitneypanye@sjtu.edu.cn

李韶旭,男,博士研究生,主要研究方向为风格化三维人脸生成。E-mail: lishaoxu94@163.com

谭帅,男,博士研究生,主要研究方向为数字人驱动与生成。E-mail: tanshuai0219@sjtu.edu.cn

韦俊杰,男,硕士研究生,主要研究方向为风格化三维人脸生成与交互。E-mail: danbaiwei@163.com

翟广涛,男,教授,主要研究方向为多媒体信号处理。E-mail: zhaiguangtao@sjtu.edu.cn

杨小康,男,教授,主要研究方向为视频编码与通信、图像处理与模式识别、视频分析与检索。

E-mail: xkyang@sjtu.edu.cn