语音驱动说话数字人视频生成方法综述

刘 颖[1]，李济廷[1*]，柴瑞坤[2]，位纪伟[2]，杨 阳[2]

（1. 军事科学院 军队政治工作研究院，北京 100166；2. 电子科技大学 计算机科学与工程学院，成都 611731）

摘要 近年来，深度学习技术的飞速发展极大地推动了虚拟数字人技术的进步，尤其是在说话数字人视频生成方面。该领域的研究在视频翻译、电影制作和虚拟助手等多个场景中展现出广阔的应用前景。该文对当前语音驱动说话数字人视频生成方法及研究现状进行了梳理与总结，并深入探讨了关键技术、数据集以及评估策略。在关键技术方面，生成对抗模型、扩散模型和神经辐射场等人工智能技术均发挥了重要作用。数据集的规模和多样性对于模型训练至关重要，而评估策略的完善则有助于更加客观地评价生成效果。说话数字人视频生成技术将继续面临众多挑战与机遇，期待该领域能够持续创新与发展，为人类社会带来更多便捷与乐趣。

关键词 说话数字人；视频生成；生成对抗模型；扩散模型；神经辐射场；多模态融合

中图分类号 TP391 **文献标志码** A DOI 10.12178/1001-0548.2024156

A Review on Audio-Driven Digital Human Generation Methods

LIU Ying[1], LI Jiting[1*], CHAI Ruikun[2], WEI Jiwei[2], and YANG Yang[2]

(1. Military Political Work Research Institute of the Academy of Military Sciences, Beijing 100166, China;

2. School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

Abstract In recent years, the rapid development of deep learning technology has greatly promoted the progress of virtual digital human technology, especially in the area of audio-driven digital human video generation. Research in this field has shown broad application prospects in various scenarios such as video translation, film production, and virtual assistants. The current methods and research status of audio-driven digital human video generation are sorted out and summarized in this paper, focusing on the key technologies, datasets, and evaluation strategies. In terms of key technologies, artificial intelligence technologies such as generative adversarial networks, diffusion models, and neural radiance fields have all played an important role. The scale and diversity of datasets are crucial for model training, and the improvement of evaluation strategies helps to evaluate the generation effect more objectively. The technology of audio-driven digital human video generation will continue to face numerous challenges and opportunities. It is expected that this field can continue to innovate and develop, bringing more convenience and fun to human society.

Key words audio-driven digital human; video generation; generative adversarial network; diffusion model; neural radiance field; multimodal fusion

随着人工智能技术的飞速发展，语音驱动说话数字人（Digital Human Generation, DHG）视频的生成和应用逐渐成为研究的热点领域[1]。数字人作为一种基于计算机技术的虚拟实体，在虚拟世界中模拟人类行为和交互的同时，也在现实世界中展现出广泛的应用前景。语音驱动的说话数字人是虚拟数字人领域的一个前沿研究方向，它专注于通过先进的人工智能技术，创建能够自然表达语音内容的虚拟角色。简而言之，说话数字人利用输入的音频信息，以及一张包含目标人物特征的图像或视频片段，通过信息提取、语义扩充、融合和对齐等步骤，生成一段目标人物自然表达这些信息的视频[2]。这一技术的核心在于多模态数据的融合和呈现，旨在以直观的视觉形式展现目标人物的语音

收稿日期：2024−06−30；修回日期：2024−08−29

基金项目：国家自然科学基金（62306067）

作者简介：刘颖，博士，副研究员，主要从事社交媒体内容分析方面的研究。

*通信作者 E-mail: 1005034741@qq.com

内容。

在语音驱动说话数字人生成的过程中，生成对抗模型（Generative Adversarial Network, GAN）、扩散模型（Diffusion Model, DM）和神经辐射场（Neural Radiance Field, NeRF）等生成方法发挥着至关重要的作用[3]。

本文旨在综述这 3 种生成方法在语音驱动的数字人生成领域的应用进展，以及它们所带来的应用前景和面临的挑战。

生成对抗模型[4] 以其独特的对抗性训练机制，使得生成的数字人图像在真实性和多样性上取得了显著成果。生成对抗模型在数字人生成中的应用，不仅提高了数字人的逼真度，还为数字人赋予了丰富的面部表情和姿态变化。随着生成对抗模型技术的不断发展，我们可以期待在未来看到更加细腻、生动的数字人形象，以及更加自然的交互体验。

扩散模型[5] 作为一种新型的生成模型，通过模拟扩散和逆扩散过程来生成数据，为数字人生成领域带来了新的可能性。扩散模型能够生成高质量、多样化的数字人图像，并且具有较好的鲁棒性和可控性。在数字人定制方面，扩散模型可以实现对数字人特征的精确控制，满足不同用户的需求。此外，扩散模型还可以应用于数字人动画的生成，为数字人的动态表现提供更多可能性。

神经辐射场[6] 方法则通过神经网络对三维场景进行编码和渲染，为数字人生成提供了三维化的解决方案。神经辐射场方法能够实现对数字人三维场景的精确重建和渲染，使得生成的数字人具有更加真实的三维感和立体感。在虚拟现实（Virtual Reality, VR）和增强现实（Augmented Reality, AR）领域，神经辐射场方法将使得数字人能够更好地融入虚拟世界，为用户带来沉浸式的体验。此外，神经辐射场方法还可应用于数字人的虚拟试衣、虚拟化妆等领域，为时尚产业带来全新的商业模式。

综合这 3 种生成方法的应用前景，可以预见数字人将在多个领域展现出广泛的应用潜力。在娱乐产业中，数字人将成为虚拟偶像、游戏角色等的重要载体，为用户带来全新的娱乐体验。在医疗领域，数字人可用于模拟手术过程、进行远程医疗等，为医疗教育和实践提供有力支持。在教育领域，数字人可用于虚拟课堂、在线教育等场景，为学生提供更加生动、直观的学习体验。此外，数字人还可应用于虚拟现实社交、虚拟导游和虚拟客服等领域，为人们的生活带来更多便利和乐趣。

本文关于语音驱动说话数字人生成的论述主要分为 3 个部分，包括关键技术、数据集以及评估策略。 最后将探讨说话数字人视频生成方法当前面临的挑战和未来的发展。

1 关键技术

早期的说话数字人生成的技术主要是依靠计算机图形学以及传统机器学习的技术，但是这类技术生成的视频质量很差，合成痕迹十分明显。随着深度学习的快速发展，图像、视频生成领域也有了许多突破，引起了说话数字人生成领域研究者的高度关注。最初，研究者通过使用自动编码器（Auto Encoder, AE）技术实现更逼真的说话数字人视频生成，ATVG[7] 和 MakeItTalk[8] 都是较为经典的模型，使得说话数字人生成领域在生成质量上有了显著提升。随着研究者对模型的不断优化与改进，目前主流的说话数字人生成技术主要分为以下 3 类：1）生成对抗模型；2）扩散模型；3）神经辐射场。图 1 为语音驱动说话数字人视频生成方法总览。表 1 为近年来说话数字人生成的模型发展总览。接下来，将依据以上 3 个部分介绍主流的说话数字人生成技术原理，并对相应的模型进行阐述。

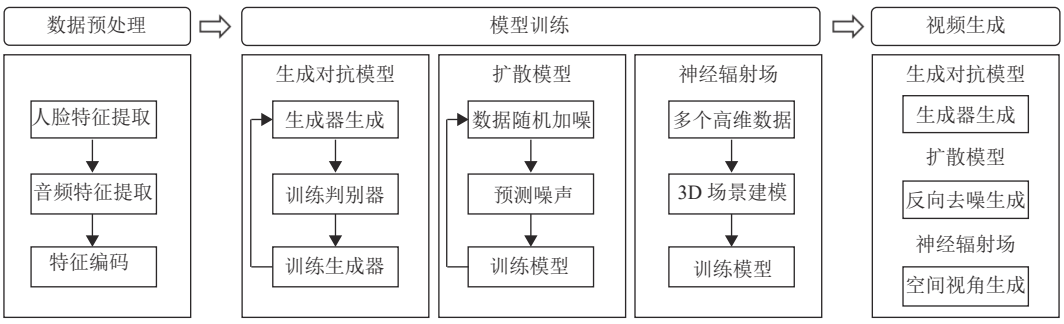

图 1 语音驱动说话数字人视频生成方法总览

表 1　说话数字人模型总览

模型	时间/年	技术框架
ATVG[7]	2019	AE
MakeItTalk[8]	2020	AE
Wave2lip[9]	2020	GAN
PC-AVS[10]	2021	GAN
AD-NeRF[11]	2021	NeRF
StyleTalker[12]	2022	GAN
video_retalking[13]	2022	GAN
SadTalker[14]	2023	GAN
GeneFace[15]	2023	NeRF
ER-NeRF[16]	2023	NeRF
DiffTalk[17]	2023	DM
DreamTalk[18]	2023	DM
EMO[19]	2024	DM
VASA-1[20]	2024	DM
GaussianTalker[21]	2024	NeRF

1.1　基于生成对抗模型的语音驱动说话数字人

生成对抗模型由两个主要部分组成：生成器（Generator）和判别器（Discriminator）[22]。其工作原理基于一种对抗过程，其中生成器尝试生成逼真的数据样本，而判别器则尝试区分这些生成的样本与真实的样本。在 GAN 中，生成器的目标是学习真实数据的分布，并生成与真实数据尽可能相似的样本。这通常是通过接收一个随机噪声向量作为输入，并输出一个与真实数据相似的样本来实现的。判别器则是一个二分类器，其目标是区分输入的数据样本是真实的还是由生成器生成的。判别器接收一个数据样本作为输入，并输出一个概率值，表示该样本是真实的概率。在训练过程中，生成器和判别器进行对抗。生成器尝试生成逼真的样本以欺骗判别器，而判别器则尝试不断提升自己的判别能力。这个过程通过交替优化生成器和判别器的参数来实现，直到达到一个平衡状态，即生成器生成的样本足以欺骗判别器，而判别器也足够准确地判断输入样本的真伪。其应用于数字人任务中的具体流程如图 2 所示，生成器负责从特征中学习生成逼真的数字人视频帧，判别器负责判断视频帧的真伪，通过多轮对抗训练，使得生成器能够生成十分逼真的数字人视频帧。

文献 [9] 首先提取音频中的关键特征，并将这些特征与相应的人脸图像进行精确配对，从而形成一个个独特的音频–图像对。随后，利用这些配对的数据来训练一个专业的音频与口型同步判别器。再引入生成对抗模型来深入探索音频–图像对之间的内在映射关系。在这个框架中，生成器网络担当着核心角色，它负责根据输入的音频特征生成高度逼真的嘴唇动作图像。与此同时，判别器网络则担任着评估者的角色，它负责判断生成的嘴唇动作与音频特征之间的同步性和真实程度。通过反复地训练和迭代，生成器网络会不断地根据判别器网络的反馈进行调整和优化。这个过程中，生成器网络会逐渐学习到如何更准确地捕捉音频中的信息，并据此生成与音频完美匹配的嘴唇动作。为了进一步探索头部姿势的操控，文献 [10] 在特征学习和图像重建的整合框架内，成功实现了对头部姿势的自由操控。这一突破的核心在于在潜空间内隐式定义了一个 12 维的姿态编码。这个编码允许在一个统一的框架中，重新定位并探索与嘴型及语音内容相关的说话内容空间，以及专门用于表达头部运动的姿态空间。通过这种设计，能够灵活调整图像中的头部姿势，同时保持与语音内容的自然同步。

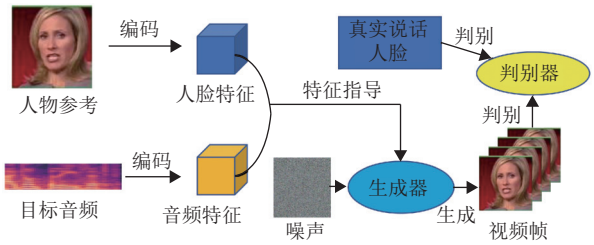

图 2　基于生成对抗模型的语音驱动说话数字人

接着研究者思考如何从单个参考图像而不是整个视频来合成具有精准音频同步唇形的说话人视频，文献 [12] 能够从单个参考图像中合成说话人的视频，该视频不仅具有精确的音频同步唇形，还展现出逼真的头部姿势和自然的眨眼动作。为实现这一目标，引入对比唇同步鉴别器，以确保唇形与音频的精准同步。此外，采用了一种条件顺序变分自动编码器，该编码器能够学习唇运动与潜在运动空间相分离的表示，从而允许独立地操纵头部运动和嘴唇运动，同时保持说话人的身份特征。为了学习复杂的音频到运动的多模态潜在空间，还运用了自回归先验增强归一化流技术。这些技术的结合能够生成高度逼真的、与音频紧密同步的说话人视频。如何增强视频的清晰度和分辨率也十分重要，文献 [13] 对用作姿势参考的随机图像帧进行了精心修改。通过中和图像中的面部表情，确保了在将其作为姿势参考输入模型时，不会带入任何嘴唇相关的特征信息，这一步骤是为了防止模型因这些图像帧中潜在的嘴唇信息而受到误导，从而避免了将这些信息作为先验知识而影响模型的训练和生成效

果。并设计了一个图像增强网络对最终生成的数字人面部尤其是嘴部进行细节增强与画质超分。

文献 [14] 则将研究进一步深入到了 3D 人脸视频的生成。为了生成具有真实感的 3D 人脸视频，首先生成 3DMM（3D Morphable Model）的三维系数，这些系数涵盖了头部姿势和表情变化。利用先进的三维面部渲染器，生成对应的视频帧。为了捕捉并学习逼真的运动场系数，深入研究了音频与不同类别运动场系数之间的内在联系，并通过模型来建模这种联系。为了进一步提升面部表情的准确性，提出了 ExpNet 模型，该模型能够精准地蒸馏运动场系数，并与三维渲染的人脸相结合，学习并生成准确的面部表情。对于头部姿势的模拟，设计了姿态变分自编码器，它能够生成具有不同风格和自然度的头部动画。这些生成的头部动画不仅符合音频的节奏，还能为视频增添丰富的动态效果。

GAN 模型具有强大的生成能力，能生成高质量的、逼真的说话人脸图像和视频。通过训练，可以学习到人脸的各种特征，包括表情、姿态和唇形等，生成与音频内容相匹配的人脸图像。它还具有一定的灵活性，可以根据输入的音频内容和人脸特征生成不同风格的说话人脸图像，也可以应对复杂的人脸场景，包括不同角度、不同姿态和不同表情。但是 GAN 的训练容易出现模式崩溃等问题，即生成的图像缺乏多样性，只包含少数几种模式。GAN 模型在训练过程中可能会过度拟合训练数据，对未见音频内容和人脸特征可能无法很好适应。

1.2 基于扩散模型的语音驱动说话数字人

扩散模型主要基于两个核心过程：前向扩散过程（Forward Diffusion Process, FDP）和反向扩散过程（Reverse Diffusion Process, RDP）[23-24]。前向扩散过程是一个参数化的马尔可夫链，它逐渐在原始数据上添加噪声，直到数据最终变为纯噪声。这个过程可以看作是对数据逐渐“模糊化”或“破坏”的过程，使得数据逐渐失去其原始特征。具体来说，给定一个初始数据分布，每一步的扩散过程都是对前一步的数据添加高斯噪声，得到新的数据。反向扩散过程是前向过程的逆过程，它从纯噪声开始，逐步去除噪声，恢复出原始数据。这个过程通过一个深度学习模型（通常是一个卷积神经网络）来实现，该模型被训练来预测前向过程中加入噪声的逆过程，即如何从噪声图像中逐步恢复出原始图像。其应用于数字人任务中的具体流程如图 3 所示，通过特征指导的说话人视频帧加噪的过程训练神经网络对噪声预测的能力，生成时通过对噪声逐步去噪的过程生成对应的数字人视频帧。

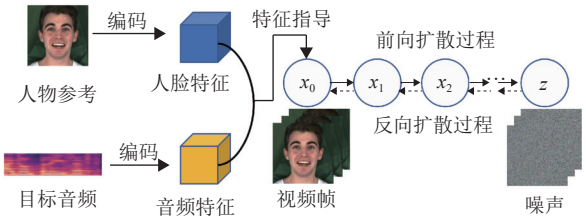

图 3 基于扩散模型的语音驱动说话数字人

文献 [17] 通过引入平滑的音频特征作为条件，成功地改进了扩散模型，使其能更有效地捕捉和模拟时序连贯的面部运动。为了进一步提升人脸建模的个性化程度，将参考人脸图像和人脸关键点作为额外的输入条件，作为模型的主要驱动因素。这一举措使得数字人的生成过程变得更为可控，并赋予了模型跨不同身份进行泛化的能力，无需任何额外的微调步骤。为了提升生成视频的质量以及稳定性，文献 [18] 利用扩散模型的优势，旨在提供卓越的性能，同时显著降低对参考风格的依赖。核心组件包括去噪网络、风格感知的嘴唇专家以及风格预测器。去噪网络巧妙地运用扩散模型，能够根据参考视频指定的说话风格，生成由音频驱动的面部动作，实现自然且流畅的语音同步。风格感知的嘴唇专家模块专注于确保嘴唇动作的精确度和表情的生动性，使得生成的数字人更加逼真且富有表现力。而风格预测器则能够直接通过音频输入，预测并生成个性化的说话风格，无需额外的风格参考。该模型在各种说话风格下均能一致地生成高度逼真的数字人，且显著减少了对参考风格的需求。不仅如此，它还具备灵活操纵说话风格的能力，能够在多语言、嘈杂音频以及领域外肖像等多种复杂输入条件下展现出强大的泛化能力。

文献 [19] 则是消除了对中间表示或繁琐预处理步骤的需求，从而极大地简化了说话头部视频的创建过程。为了提升生成过程中的稳定性，在模型中引入控制机制，包括速度控制器和面部区域控制器。这两个控制器作为可调的超参数和微调的控制信号，既增强了生成过程的稳定性，又不损害最终视频的多样性和表现力。为了确保生成的视频中角色的外观与输入参考图像保持一致，采用并强化 ReferenceNet[25] 的方法。这一方法确保了模型在生成视频时能够准确捕捉并保留角色的独特特征。该

模型不仅能制作令人信服的演讲视频，更能够灵活制作出各种风格的歌唱视频。在表现力和真实性方面，该模型均得到了显著提升。为了进一步增强生成的速率与泛用性，文献 [20] 训练了一个扩散 Transformer 模型，专注于整体面部动力学和头部运动的潜在空间。在此过程中，将面部动力学的所有可能表现形式——包括嘴唇的运动、非嘴唇区域的表情变化、眼睛凝视方向以及眨眼等——整合为一个统一的潜在变量，以系统的方式对其概率分布进行建模。为了进一步增强模型的灵活性和适应性，引入一组可选的条件信号，这些信号涵盖了主凝视方向、头部与观察者的距离以及情绪偏移等关键信息。该模型不仅能够生成与音频精确同步的嘴唇动作，而且能够捕捉到丰富的面部细微变化以及自然的头部运动。这些特性极大地提升了生成内容的真实感和生动性。

扩散模型通常能够生成高质量的图像和视频，因为它们基于概率扩散过程，可以逐步从噪声中恢复出目标图像，从而保留更多的细节和纹理，同时具有相对较高的灵活性，可以适应不同的数据集和任务需求。但是扩散模型的训练和推理过程通常需要较高的计算成本，因为需要进行多次迭代来逐步从噪声中恢复出目标图像，这也导致其生成速度较慢，在需要快速生成结果的应用场景中，表现不如其他生成模型[25]。

1.3　基于神经辐射场的语音驱动说话数字人

神经辐射场是一种基于深度学习的三维场景渲染方法，其核心原理是利用神经网络来表示和渲染复杂的三维场景[26-27]。神经辐射场通过神经网络学习场景的隐式表示，即不直接表示场景的几何形状和纹理，而是通过神经网络学习出一个连续的三维函数，该函数可以给出空间中任意位置和方向上的颜色和密度。神经辐射场的输入包括相机姿态（即相机的位置和朝向）和真实图像，输出则是场景的隐式表示。具体来说，对于每个空间点，神经辐射场会考虑其 3D 坐标位置 (x,y,z) 以及观察方向 (θ,ϕ)，然后输出该点的颜色和密度。神经辐射场采用体渲染（Volume Rendering, VRe）的方法来进行渲染。体渲染是一种通过计算光线穿过三维场景时与场景中物质交互产生的颜色和亮度来生成图像的技术。在神经辐射场中，对于从相机发出的每条射线，神经辐射场会计算射线上每个点的颜色和密度，并将它们进行加权求和，从而得到该射线的最终颜色，其具体流程如图 4 所示。

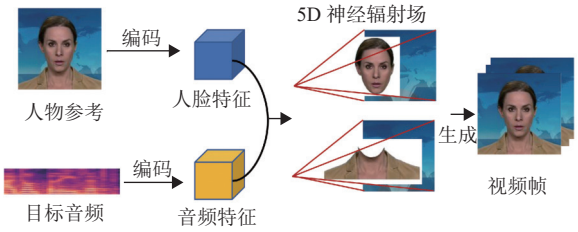

图 4　基于神经辐射场的语音驱动说话数字人

文献 [11] 通过将音频特征直接输入到一个约束隐式函数中，构建了一个能够生成动态神经辐射场的模型。随后利用体渲染技术，合成高保真度的说话人脸视频。采用了两个独立的神经辐射场：一个专注于头部区域的渲染，而另一个则负责身体部分的生成。这种设计使得模型具备了强大的视频编辑能力，不仅可以轻松编辑人物的姿态，还能替换背景。文献 [15] 为了克服传统基于神经辐射场方法在泛化能力方面的局限性和在处理人脸数据时常见的人脸均值问题，提出了利用大语料库训练一个变分运动生成器，以构建一个通用的音频到运动的映射模型。引入一个域自适应后网络，核心功能是将通过音频预测得到的运动表示有效地整合到特定的目标人域中，从而实现个性化的运动映射。

文献 [16] 则进一步探索了条件神经辐射场在实时渲染和快速收敛方面的潜力。基于条件神经辐射场，能够在较小的参数量下同时达到高精度的实时渲染和快速收敛的效果。通过设计 3 个 2D 哈希编码器，对空间区域进行了智能修剪，并引入一种紧凑而强大的三平面哈希表示，这种表示基于神经辐射场，极大地提升了场景渲染的效率和表现力。针对语音音频，提出了区域注意模块。该模块利用区域注意力机制，生成了区域感知的条件特征，从而能够更准确地捕捉与音频相对应的局部运动信息。这种注意力机制在音频特征和空间区域之间建立了直接的显式连接，提供了关于局部运动的先验知识。对于身体部分，提出了一种直观且高效的适应性姿态编码方法，将头部姿态的复杂变换精确地映射到空间坐标中，从而显著优化了头部与躯干之间的分离问题，使得生成的动画更加自然和流畅。

文献 [21] 在三维高斯散射（3D Gaussian Splatting, 3DGS）的基础上，提出了一种新的面部运动控制方法。利用 3D 高斯分布的明确数学表达，将其巧妙地与 3D 面部模型相结合，从而实现对面部运动直观且精准的控制。该模型由两大核心

组件构成：特定人物运动翻译器和动态高斯渲染器。特定人物运动翻译器首先提取通用音频特征，然后结合自定义的嘴唇运动生成技术，确保能够精确捕捉并重现目标说话者的独特嘴唇动作。不仅提高了系统的个性化程度，也极大增强了其模拟真实人物面部运动的准确性。动态高斯渲染器引入了特定于说话者混合形状技术，通过结合潜在姿势，能够显著提升面部细节的表现力，使得渲染出的视频既稳定又逼真。不仅使面部动画更加自然流畅，也进一步提升了整个系统的实用性和用户体验。

NeRF 能够生成高质量、逼真度高的三维模型。这意味着在任意角度和距离下，它都能呈现出真实的物体表面和纹理细节，对于人脸这样的复杂对象尤为重要。同时通过使用两个独立的神经辐射场（一个用于头部区域，一个用于身体部分），NeRF 模型具备强大的视频编辑能力，包括改变人物的姿态、替换背景等。而且 NeRF 可以从任意数量的输入图像中生成三维模型，不需要对输入进行特定处理或标记，这使得它在实际应用中具有更高的灵活性和适用性。但是 NeRF 技术需要大量的计算资源和时间进行训练，在处理大规模场景和复杂光照条件时可能会遇到困难，影响到生成视频的质量和真实性。由于 NeRF 是基于视图合成的技术，因此在生成模型时需要确保输入的视角足够广泛和充分。如果输入视角不足，可能会导致模型中存在遮挡和空洞。目前大多数基于 NeRF 的方法在融合音频和视觉特征时缺乏有效的监督式多模态特征融合方法，有时可能无法精准地将音频映射到与语音运动相关的面部区域。

这 3 类主流的说话数字人生成技术使得生成的数字人更逼真和谐，面对不同场景，这 3 类方法有着不同的优劣，如表 2 所示。

表 2　语音驱动说话数字人方法综合比较

模型	场景	优势	劣势
GAN	通用人	画面质量高	训练不稳定
DM	通用人	细节丰富	计算资源需求高
NeRF	特定人	真实感强	计算资源需求高

2　数据集

数据集在说话数字人生成任务中扮演着重要的角色。在说话数字人生成任务中，深度学习作为一种典型数据驱动技术，需要充足的数据来训练模型才能使模型能够学习到该任务的特点。一方面，数据集可以充当评估不同说话数字人生成算法性能的通用平台；另一方面，数据集的规模和多样性也给该任务带来了日益复杂的挑战。其中数据集规模的增大，深度学习模型可以学习到数据中更为复杂的特征；高分辨率数据集则使得深度学习模型可以生成细节丰富且更为逼真的说话数字人；此外，更加多样性的数据集（包括表情、头部动作和身体姿势等）能够帮助模型更好地泛化，使得生成的说话数字人更能满足现实生活的需要。根据应用场景和目标的不同，将数据集划分为通用说话数字人数据集和特定说话数字人数据集，本章主要介绍这两类中具有代表性的数据集，其整体情况如表 3 所示，其中“-”表示不适用。

表 3　常用语音驱动说话数字人生成数据集

数据集	时间/年	人数	有无明显头部动作	有无情感标签	是否预处理	类别
GRID[28]	2006	33	无	无	-	通用
CREMA-D[29]	2014	91	无	有	-	通用
LRW[30]	2017	1 000+	无	无	-	通用
ObamaSet[31]	2017	1	有	无	-	特定
MEAD[32]	2020	60	有	有	-	通用
HDTF[33]	2021	362	有	无	-	通用
May[34]	2023	1	-	-	是	特定
PXB184[35]	2024	1	-	-	是	特定
RLW104[35]	2024	1	-	-	是	特定
TXB805[35]	2024	1	-	-	是	特定
GQS883[35]	2024	1	-	-	是	特定

2.1　通用说话数字人数据集

通用说话数字人数据集指的是包含多种不同人物、语音和面部运动数据的数据集。这类数据集通常具有广泛的适用性和泛化能力，因为它们覆盖了不同的人物特征、语音风格和面部表情。

在 GRID 数据集[28]（The QMUL Underground

Re-Identification Dataset）中，包含 34 名说话个体（18 名男性和 16 名女性），每个个体都直面相机且分别读出 1 000 条短句，这些短句都是由一个仅包含 51 个单词的小型字典中随机选出的 6 个单词组成。其中所有个体在说话时都没有明显的情绪波动和头部动作。该数据集由英国萨里大学于 2006 年发布，通过招募志愿者到安静的录音室内进行数据采集，以确保获得高质量的音频和视频，此外对录制的音频和视频也进行了一定的整理，去除其中的冗余部分以消除误差。

在 CREMA-D 数据集[29]（Crowd-Sourced Emotional Multimodal Actors Dataset）中，包含年龄段在 20～74 岁的 91 名演员（48 名男性和 43 名女性），不同于其他数据集，每个演员都分别以不同类别的情感（总共包含 6 类情感）和情感强度（总共包含 4 种情感强度）来重复说出同一个句子，每人总共说出 12 条句子且在说话时没有明显的头部动作。该数据集是专业演员在专业导演的监督下录制而成，其中由于演员受到过专业训练，能够以不同类别的情感和情感强度传达目标情绪，而导演则是通过描述专门设计的场景来引导演员唤起目标情绪。

在 LRW 数据集[30]（The Lip Reading in the WildDataset）中包含了 500 个单词的视频片段，由数百个说话个体读出，其中一个比较特殊的地方是：一些视频仅包含一个直面相机的说话个体，而另外的视频则包含不止一名说话个体的小组辩论。该数据集是从 BBC 的电视广播中收集得到，每个视频片段较短，且个体说话时没有明显的头部动作。该数据集是由英国伦敦帝国理工学院创建的，其中的视频来源于新闻、访谈和脱口秀等多种类型节目，在原视频上进行截取以获得所需的视频片段，且每个视频片段都经过手动检查和验证，确保包含正确的单词和唇部运动。

MEAD（Multi-view Emotional Audio-Visual Dataset）是一个大规模音视频情感说话人脸数据集[32]，专门用于生成带有特定情感的说话人脸。该数据集中包括 60 名演员（30 名男性和 30 名女性），每位演员以不同类别的情感（总共包含 8 种情感）和情感强度（总共包含 3 种情感强度）在严格控制的环境中从多个角度录制说话视频，以确保捕捉到演员说话时的面部表情细节。该数据集由商汤科技等机构联合发布，在视频录制前，由相关人员辅助演员进入指定的情感状态，并在录制时要求演员不能出现停顿或发音错误等问题，录制完成后仍需指导团队来评判是否符合要求，通常每个视频片段都需要演员录制两到三次以确保视频质量。

HDTF（aHigh-resolution Audio-visual Dataset）是一个高分辨率的说话数字人数据集[33]，该数据集从 YouTube 上获取，并经过处理和标注，用于高分辨率说话数字人生成。HDTF 数据集由大约 362 个不同的视频组成，总时长为 15.8 h，且视频分辨率都为 720 P 或 1 080 P。该数据集视频片段来源于访谈、演讲、电视剧和电影等多种类型节目，以确保视频源的多样性，能够涵盖不同性别、年龄和种族的说话人，其中视频分辨率都较高以确保视频中人物的面部区域清晰可见并且光照良好。

2.2　特定说话数字人数据集

特定说话数字人数据集专注于一个或少数几个特定人物的语音和面部运动数据。这类数据集通常用于生成特定人物的逼真虚拟形象，如虚拟偶像、电影角色或历史人物等。

ObamaSet 是一个特定的视听数据集[31]，专注于分析美国前总统奥巴马的演讲，为特定说话数字人生成任务提供了数据基础。该数据集中的所有视频都来自于奥巴马的每周演讲，与以往的数据集不同，ObamaSet 只包含一个说话主体并且不提供其他人工注解。

May 是一个预先处理好的特定人物数据集[34]，在其原有视频数据的基础上进行了裁剪、音频重采样等操作。目前主流使用 NeRF 来生成说话数字人，都是收集特定人物的视频数据集，然后在该基础上进行一定的预处理操作并形成相应的 NPZ 文件（NumPy 库中用于存储多个 NumPy 数组的标准压缩文件格式），而后再通过特定的渲染器即可得到该人物的 NeRF 模型。

PXB184、RLW104、TXB805 和 GQS883 都是由 Meta AI 团队收集并制作的人物模型数据集[35]。Meta AI 团队首先收集了一组丰富的双人对话数据集，而后从这些数据集构建复合运动模型，其中包括特定人物的面部、姿势和身体运动模型，之后通过加载数据集中的模型参数等，可以实现音频驱动的特定说话数字人生成。

3　评估策略

在说话数字人生成领域中，如何衡量一个说话数字人生成模型的效果也是相当重要的。常用的说话数字人生成评估策略主要分为主观评估和客观评估两大类，但二者都有着一定的局限性。如主观评

价中，由于个体的差异性，用户测试是不可重复且不稳定的；又或者是客观评价中，虽然已经有了很多关于各方面的评估指标，但是不同指标之间可能是不适配，甚至是相互矛盾的。因此一般同时采用主观评估和客观评估来全方位地评估一个说话数字人生成模型。下面将分别对这两类评估方法进行介绍。

3.1 主观评估策略

主观评估是说话数字人生成任务中最直观的评估方法，由于生成的说话数字人是面向现实世界并展现给用户使用的，因此通过这种评估方式能够明显地展现出哪些地方有较大的不足。评估的方式一般可以分为定性评价和用户测试。其中定性评价一般取自真实视频的一些视频帧，然后展现各个说话数字人生成方法根据相应输入而后输出的对应视频帧片段，通过对比这些视频帧，可以直观地比较说话数字人生成视频的视觉质量、人物是否保留原有特征等，若是某些方法可以改变说话数字人的表情或艺术风格等，还可以对生成效果进行判断。通过定性评估可以很直观地看出各个说话数字人生成方法的优缺点、生成视频质量等，不过考虑到过拟合等情形，可能会出现某些方法在特定人物视频上展现出很不错的效果，在总体性能上的表现却没有多好，从而会对评估结果造成一定的影响，因此该方法仍然存在一定的缺点。

而用户测试一般是同比例从男性和女性中选出测试者，组成一定规模的测试者群体，然后将各种模型生成的说话数字人视频随机展现给这些测试者观看，确保测试者事先不知道每个说话数字人视频是由哪个模型生成的，然后由这些测试者对这些视频进行评分，通过收集用户测试的结果来评估各个说话数字人模型在视频真实性、头部动作以及音频一致性等方面的效果。由于个体的差异性，使用该方法得到的评估结果可能会造成一定的偏差。为了消除偏差，一方面可以挑选更多的测试者，使得测试结果能够代表更广泛的用户；另一方面则是可以结合多种更为客观的评价指标来共同评估模型效果。

3.2 客观评估策略

对于生成的说话数字人视频，我们可以从多个方面来进行评估，比如生成视频的清晰度等视觉质量、语音与说话数字人口型是否对齐以及说话数字人是否有头部动作等，表 4 展示了各类客观评估指标。下面将分别介绍主要的客观评估指标。

1）峰值信噪比（Peak Signal to Noise Ratio, PSNR）[36] 是图像处理和视频处理领域中常用的一种客观评价指标，在说话数字人生成中通过衡量原始图像与生成视频帧之间的误差来评估生成视频帧质量。具体而言，PSNR 值越大，表示生成视频帧的质量越高。假设生成的视频帧为I、原始图像为K，则 PSNR 定义为：

$$\mathrm{PSNR} = 10\log\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right) \tag{1}$$

式中，MAX表示图像上像素点的最大数值；MSE定义为：

$$\mathrm{MSE} = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\| I(i,j) - K(i,j)\|^2 \tag{2}$$

PSNR 在说话数字人评估中应用很广泛，但由于 PSNR 会依据MAX计算，而不同的图片可能会出现内容不同，但是MAX相同的情况，因此会对最终的结果造成一定的影响。

表 4 主要客观评估指标

定量评估指标	帧质量（越大越好(↑)/越小越好(↓)）
PSNR[36]	↑
SSIM[37]	↑
FID[38]	↓
LMD[39]	↓
SyncNet score[40]	↑
LSE-C[9]	↑
LSE-D[9]	↓

2）结构相似度（Structural Similarity Index, SSIM）[37] 是图像质量评价的一个重要指标，用于衡量两幅图像之间的相似度。与 PSNR 不同，SSIM 不仅考虑图像的像素差异，还考虑图像的结构信息、对比度和亮度。因此，SSIM 的评估效果能够更好地反映人眼对图像质量的感知效果。此外，SSIM 在图像的局部质量的评估上表现也相对较好，而这也符合现实生活中人们对一张图片的直观观察，即关注于图像中比较重要的部位，如人物图像中主要观察其中的人物形象而对周围的背景环境进行一定的忽视。

3）FID[38] 用于评估生成的人脸图像与真实人脸图像之间的相似性。通常通过一个预训练的深度神经网络来提取图像特征，并利用这些特征计算生成图像和真实图像之间的 Fréchet 距离，具体而言，FID 分数越低，表示生成的人脸图像越接近真

实人脸图像。定义如下：

$$\mathrm{FID}=\| u_{\mathrm{r}}-u_{\mathrm{g}}\|_2^2+\mathrm{Tr}\left(\boldsymbol{\Sigma}_{\mathrm{r}}+\boldsymbol{\Sigma}_{\mathrm{g}}-2\left(\boldsymbol{\Sigma}_{\mathrm{r}}\boldsymbol{\Sigma}_{\mathrm{g}}\right)^{\frac{1}{2}}\right) \tag{3}$$

式中，μ_{r}和μ_{g}分别是真实图像和生成图像特征分布的均值；Tr表示矩阵的迹；$\boldsymbol{\Sigma}_{\mathrm{r}}$和$\boldsymbol{\Sigma}_{\mathrm{g}}$分别是真实图像和生成图像特征分布的协方差矩阵。

4）关键点距离（LMD）[39] 用于评估面部图像生成质量，可以用于度量生成人脸图像与真实人脸图像之间嘴部特征点的距离，以评估生成人脸图像中人物嘴型的准确性。定义如下：

$$\mathrm{LMD}=\frac{1}{N}\sum_{i=1}^{N}\sqrt{(x_{\mathrm{g}_i}-x_{\mathrm{r}_i})^2+(y_{\mathrm{g}_i}-y_{\mathrm{r}_i})^2} \tag{4}$$

式中，N是特征点的数量；x_{g_i}和y_{g_i}分别是生成图像上第i个特征点的x和y坐标；x_{r_i}和y_{r_i}分别是真实图像上第i个特征点的x和y坐标。LMD 除了可以用于计算人物嘴型的准确性，还可以对面部的其他位置进行评估。

5）唇音一致性得分（SyncNet Score）[40] 用于衡量生成的说话数字人视频与音频之间的一致性问题，分别通过 SyncNet 网络中视频和音频的预训练模块来提取视频特征和音频特征，计算二者之间的距离，再根据距离来计算得分 S，得分越高则表示生成的说话数字人视频与音频越同步。定义为：

$$S=M(d)-\min(d) \tag{5}$$

式中，$d=\| f_{\mathrm{a}}-f_{\mathrm{v}}\|_2^2$，$f_{\mathrm{a}}$和$f_{\mathrm{v}}$分别表示神经网络提取的视频特征和音频特征；$M$ 表示中位数。

6）距离评分（LSE-C）和置信评分（LSE-D）[10] 二者均基于 SyncNet 网络，分别用来评估说话数字人的口型以及唇部同步性能。LSE-C 通常通过余弦相似度来计算音频和视频特征之间的相似度，定义如下：

$$\mathrm{LSE-C}=\frac{1}{N}\sum_{i=1}^{N}\mathrm{Cosine\ Similiarity}(f_{\mathrm{a}_i},f_{\mathrm{v}_i}) \tag{6}$$

式中，f_{a_i}和f_{v_i}分别表示第i帧的音频特征和视频特征。而 LSE-D 通常通过欧式距离计算音频和视频特征之间的距离，定义为：

$$\mathrm{LSE-D}=\frac{1}{N}\sum_{i=1}^{N}\| f_{\mathrm{a}_i}-f_{\mathrm{v}_i}\|_2 \tag{7}$$

总体而言，LSE-C 值越大、LSE-D 值越小表示生成的说话人视频效果更好以及具有更好的同步性。

4　未来重点展开工作

本文从关键技术、数据集和评估策略 3 个方面总结了说话数字人生成技术当前的现状。总体而言，受益于以深度学习为基础的人工智能技术的迅猛发展，当前的说话数字人视频生成技术已取得了显著进展，但仍然面临诸多挑战。

4.1　提升生成视频细节质量

说话数字人视频生成中，影响视频质量的因素较多，但当前的算法仅考虑了一些主要特征，如嘴型和头部动作，导致生成结果缺乏表现力，人物的动作、表情和眼神显得呆板，缺乏生动感。人类对合成视频中的任何运动变化都非常敏感，他们会无意识地关注嘴唇、眼睛、眉毛和自发的说话运动。具有视听一致性的唇部运动对于说话数字人视频生成十分重要，然而当前的研究对隐性特征，如眼睛和情感特征等的关注度还不够高。要解决这些问题，一方面需要更加注重细节的生成，包括眼睛和情感等隐性特征的生成；另一方面，需要开发新的训练方法。同时，可以构建更多具有语义和情感动作注释等隐藏特征的高质量视觉−语音数据集。

4.2　降低计算复杂度

当前大多数模型的生成速度过慢，特别是目前基于扩散的方法，许多工作依赖于逐帧视频生成的方式，导致时间复杂度较高。人机交互作为虚拟人类未来发展的一种方法，要求数字人具备多模态信息，如自然语言、面部表情和自然的类人手势。同时，数字人需要在接收语音等请求后迅速提供高质量的视频反馈。然而，由于大多数模型推理的时间复杂度过高，这严重限制了其在人机交互领域的应用。在人机交互中，及时响应对于改善用户体验至关重要。尽管语音模块和对话模块已经在商业应用中得到广泛使用，可以满足人机交互的实时需求，但目前说话数字人视频生成模型需要较长时间来渲染和输出多模态模型。因此，需要提高说话数字人视频生成模型的数据处理效率，减少多模态视频的渲染时间，并缩短人机交互系统的响应时间。尽管一些虚拟人类产品已实现低延迟响应，但生产周期长、成本高和便携性差等问题仍需引起重视。

4.3　构建高质量数据集

在人工智能时代，数据集是深度学习模型的重要组成部分。同时，数据集也推动了虚拟人合成领域中复杂问题的解决。然而，在实际应用中，很少有高质量的注释数据集能够满足说话数字人合成模

型的训练需求。此外，许多机构和研究人员受到深度伪造技术伦理问题的影响，使得部分数据集的获取难度增加。只有少数数据集是完全开源的，有些需要通过申请程序获得。有些只有来自大学、研究机构和企业的研究人员、教师和工程师才能申请。并且，当前可用数据集在质量和数量上都有待提升，实验室收集的数据受成本限制且规模有限，而社交媒体上收集的数据集虽然解决了数量问题，但质量良莠不齐，同时缺乏细粒度的标注。未来的数字人数据集不仅需要扩大参与个体的数量，还应注重高质量、细粒度的标注。探索其他有效的学习范式也不失为一个好的方向，如知识蒸馏和少样本学习，以挖掘说话数字人视频生成任务的潜力。

4.4 增加全面评估指标

正如前文所述，语音驱动说话数字人视频生成的评估任务是一个开放性问题，需要从客观和主观两个方面对生成结果进行评估。主观评估由选出的测试者进行评估，往往会存在显著的个体差异，导致不同测试者和方法得出的结果差异较大，缺乏稳定性与可重复性。人们对生成的数字人的准确性和真实性有很高的标准，即使轻微的缺陷也可能被视为显著的不真实，这给客观评估指标带来了极大的困难。目前的客观评估指标，如 PSNR、SSIM 等，一方面无法很好地解释人类的感知，另一方面更多局限于图像层面的视频质量评估，缺乏对视频内容实际表现如视频连贯性、人物表情动作自然性的有效客观评估方式。

5 结束语

本文深入探讨了语音驱动说话数字人视频生成方法的现状与发展。随着技术的不断进步，从生成对抗模型、扩散模型和神经辐射场等生成方法，我们见证了数字人逼真度和自然度的显著提升。数据集的不断丰富和扩展，为数字人生成提供了更为广泛和丰富的素材。而评估策略的完善，则为数字人生成的质量提供了更为客观和科学的评价标准。在娱乐、医疗和教育等行业中，说话数字人将成为重要的辅助工具，为用户带来更加生动、直观和沉浸式的体验。同时，随着技术的不断进步和应用场景的不断拓展，数字人生成技术也面临着更多的挑战和机遇。

参 考 文 献

[1] 胡青, 刘本永. 基于卷积神经网络的说话人识别算法[J]. 计算机应用, 2016, 36(A01): 79-81.
HU Q, LIU B Y. Speaker recognition based on convolutional neural network[J]. Journal of Computer Applications, 2016, 36(A01): 79-81.

[2] 宋一飞, 张炜, 陈智能, 等. 数字说话人视频生成综述[J]. 计算机辅助设计与图形学学报, 2023, 35(10): 1457-1468.
SONG Y F, ZHANG W, CHEN Z N, et al. A survey on talking head generation[J]. Journal of Computer-Aided Design & Computer Graphics, 2023, 35(10): 1457-1468.

[3] ZHEN R, SONG W C, HE Q, et al. Human-computer interaction system: A survey of talking-head generation[J]. Electronics, 2023, DOI: 10.3390/electronics12010218.

[4] CRESWELL A, WHITE T, DUMOULIN V, et al. Generative adversarial networks: An overview[J]. IEEE Signal Processing Magazine, 2018, 35(1): 53-65.

[5] STYPUŁKOWSKI M, VOUGIOUKAS K, HE S, et al. Diffused heads: Diffusion models beat GANs on talking-face generation[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. New York: IEEE, 2024: 5089-5098.

[6] LI W C, ZHANG L H, WANG D, et al. One-shot high-fidelity talking-head synthesis with deformable neural radiance field[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2023: 17969-17978.

[7] CHEN L L, MADDOX R K, DUAN Z Y, et al. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2019: 7824-7833.

[8] ZHOU Y, HAN X T, SHECHTMAN E, et al. MakeltTalk[J]. ACM Transactions on Graphics, 2020, 39(6): 1-15.

[9] PRAJWAL K R, MUKHOPADHYAY R, NAMBOODIRI V P, et al. A lip sync expert is all you need for speech to lip generation in the wild[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM, 2020: 484-492.

[10] ZHOU H, SUN Y S, WU W, et al. Pose-controllable talking face generation by implicitly modularized audio-visual representation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2021: 4174-4184.

[11] GUO Y D, CHEN K Y, LIANG S, et al. AD-NeRF: Audio driven neural radiance fields for talking head synthesis[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2021: 5764-5774.

[12] MIN D C, SONG M, KO E, et al. StyleTalker: One-shot style-based audio-driven talking head video generation [EB/OL]. [2024-01-22]. http://arxiv.org/abs/2208.10922v2.

[13] CHENG K, CUN X D, ZHANG Y, et al. VideoReTalking: Audio-based lip synchronization for talking head video editing in the wild[C]//Proceedings of the SIGGRAPH Asia 2022 Conference Papers. New York: ACM, 2022: 1-9.

[14] ZHANG W X, CUN X D, WANG X, et al. SadTalker:

Learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2023: 8652-8661.

[15] YE Z H, JIANG Z Y, REN Y, et al. GeneFace: Generalized and high-fidelity audio-driven 3D talking face synthesis[EB/OL]. [2024-02-23]. http://arxiv.org/abs/2301.13430v1.

[16] LI J H, ZHANG J W, BAI X, et al. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2023: 7534-7544.

[17] SHEN S, ZHAO W L, MENG Z B, et al. DiffTalk: Crafting diffusion models for generalized audio-driven portraits animation[EB/OL]. [2023-12-22]. http://arxiv.org/abs/2301.03786v2.

[18] MA Y F, ZHANG S W, WANG J Y, et al. DreamTalk: When emotional talking head generation meets diffusion probabilistic models[EB/OL]. [2023-12-22]. http://arxiv.org/abs/2312.09767v3.

[19] TIAN L R, WANG Q, ZHANG B, et al. EMO: Emote portrait alive: Generating expressive portrait videos with Audio2Video diffusion model under weak conditions [EB/OL]. [2024-03-10]. http://arxiv.org/abs/2402.17485v3.

[20] XU S C, CHEN G J, GUO Y X, et al. VASA-1: Lifelike audio-driven talking faces generated in real time[EB/OL]. [2024-03-10]. http://arxiv.org/abs/2404.10667v1.

[21] YU H Y, QU Z, YU Q H, et al. GaussianTalker: Speaker-specific talking head synthesis via 3D Gaussian splatting[EB/OL]. [2024-03-10]. http://arxiv.org/abs/2404.14037v3.

[22] KARRAS T, LAINE S, AITTALA M, et al. Analyzing and improving the image quality of StyleGAN[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 8107-8116.

[23] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2022: 10674-10685.

[24] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.

[25] HU L, GAO X, ZHANG P, et al. Animate anyone: Consistent and controllable image-to-video synthesis for character animation[EB/OL]. [2024-03-10]. http://arxiv.org/abs/2311.17117v3.

[26] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. NeRF: Representing scenes as neural radiance fields for view synthesis[C]//European Conference on Computer Vision. Cham: Springer, 2020: 405-421.

[27] GAO K, GAO Y N, HE H J, et al. NeRF: Neural radiance field in 3D vision, A comprehensive review[EB/OL]. [2024-01-21]. http://arxiv.org/abs/2210.00379v5.

[28] COOKE M, BARKER J, CUNNINGHAM S, et al. An audio-visual corpus for speech perception and automatic speech recognition[J]. The Journal of the Acoustical Society of America, 2006, 120(5 Pt 1): 2421-2424.

[29] CAO H W, COOPER D G, KEUTMANN M K, et al. CREMA-D: Crowd-sourced emotional multimodal actors dataset[J]. IEEE Transactions on Affective Computing, 2014, 5(4): 377-390.

[30] CHUNG J S, ZISSERMAN A. Lip reading in the wild[C]//Asian Conference on Computer Vision. Cham: Springer, 2017: 87-103.

[31] SUWAJANAKORN S, SEITZ S M, KEMELMACHER-SHLIZERMAN I. Synthesizing Obama[J]. ACM Transactions on Graphics, 2017, 36(4): 1-13.

[32] WANG K, WU Q, SONG L, et al. MEAD: A large-scale audio-visual dataset for emotional talking-face generation[C]//European Conference on Computer Vision. Cham: Springer, 2020: 700-717.

[33] ZHANG Z M, LI L C, DING Y, et al. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2021: 3660-3669.

[34] YE Z H, HE J Z, JIANG Z Y, et al. GeneFace++: Generalized and stable real-time audio-driven 3D talking face generation[EB/OL]. [2024-01-22]. http://arxiv.org/abs/2305.00787v1.

[35] NG E, ROMERO J, BAGAUTDINOV T, et al. From audio to photoreal embodiment: Synthesizing humans in conversations[EB/OL]. [2024-01-22]. http://arxiv.org/abs/2401.01885v1.

[36] HUYNH-THU Q, GHANBARI M. Scope of validity of PSNR in image/video quality assessment[J]. Electronics Letters, 2008, 44(13): 800.

[37] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: From error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[38] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. GANs trained by a two time-scale update rule converge to a local Nash equilibrium[EB/OL]. [2024-01-22]. http://arxiv.org/abs/1706.08500v6.

[39] CHEN L L, LI Z H, MADDOX R K, et al. Lip movements generation at a glance[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018.

[40] CHUNG J S, ZISSERMAN A. Out of time: Automated lip sync in the wild[C]//Asian Conference on Computer Vision. Cham: Springer, 2017: 251-263.

编　辑　税　红