# 数字说话人视频生成综述

宋一飞 [1)], 张炜 [2)], 陈智能 [1)*], 姜育刚 [1)]

[1)] (复旦大学计算机科学技术学院 中国上海 200438)

[2)] (京东探索研究院 中国北京 100101)

(zhinchen@fudan.edu.cn)

摘 要: 近年来, 基于深度学习的生成技术显著推动了虚拟数字人技术的发展. 本文针对当前虚拟数字人研究中的热点问题—数字说话人视频生成进行综述, 其在电影配音, 动画制作, 虚拟助手等场景中具有重要的应用前景. 本文从数据集、关键技术、评估策略三个方面, 对当前数字说话人视频生成技术及研究现状做一个较系统的梳理与总结, 介绍了其生成过程中涉及的视觉生成, 图像识别, 语音识别, 跨模态分析等多项人工智能的关键技术机器发展演进过程. 从数据, 模型, 评估策略等方面指出该方向需要迫切解决的问题, 并通过这些问题对其未来的发展方向作了展望, 以期能对该领域的研究者有所帮助启发, 促进该方向的发展.

关键词: 虚拟数字人;数字说话人;视频生成;多模态融合;深度学习
中图法分类号: TP391. 41 **DOI:** 10. 3724/SP. J. 1089. 2023. 19782



# A Survey on Talking Head Generation

Song Yifei[1)], Zhang Wei[2)], Chen Zhineng[1)*], Jiang Yu-Gang[1)]

[1)] (*School of Computer Science, Fudan University, Shanghai, China* 200438)

[2)] (*JD Explore Academy, Beijing, China* 100101)

**Abstract:** In recent years, the advancement of virtual digital human technology has been significantly accelerated by deep learning-based generative techniques. This paper offers a comprehensive review of the current hot topic in virtual digital human research: talking head generation. It emphasizes the promising applications of this technology in domains s such as film dubbing, animation production, and virtual assistants. From the perspectives of dataset availability, key technologies, and evaluation strategies, this paper presents a systematic overview and summary of the current state of talking head generation technology and research. It introduces pivotal artificial intelligence technologies involved in the generation process, encompassing visual generation, image recognition, speech recognition, and cross-modal analysis, along with their progressive developments. The paper identifies pressing issues that requires attention in this field, such as data, models, and evaluation metrics, and offers a future outlook based on these challenges. Its objective is to provide insightful guidance and promote the advancement of this field for researchers in the domain.

**Key words:** Virtual digital human; Talking head generation; Video generation; Multi-modal fusion; Deep learning

收稿日期: 20**-**-**; 修回日期: 20**-**-**. **基金项目**: 国家自然科学基金(62032006, 62172103); **宋一飞**(1999—) , 男, 硕士研究生, CSIG 学生会员, 主要研究方向为多媒体生成; 张炜(1986—), 男, 博士, 高级研究员, CCF 会员, 主要研究方向为计算机视觉及多媒体; 陈智能(1982—) , 男, 博士, 研究员, 博士生导师, CCF 会员, 论文通信作者, 主要研究方向为多媒体信息处理、人工智能; 姜育刚(1981—) , 男, 博士, 教授, 博士生导师, CCF 会员, 主要研究方向为计算机视觉、可信人工智能.

## 1 简介

近年来, 随着以深度学习为核心的人工智能技术的飞速发展, 虚拟数字人技术受到了越来越多的关注. 虚拟数字人技术包括人物形象的设计建模, 人脸的3D生成, 人物动作表情的捕捉与生成, 音视频的合成显示与交互等, 其特别是聚焦人物头部的虚拟数字说话人在虚拟世界中有着广阔的应用前景. 当前高质量虚拟数字人, 例如数字说话人生成主要依赖于真人的头部与面部动作捕捉, 例如高分辨率相机. 这种方法高度依赖昂贵的设备和专业的表演人员. 如何利用前沿人工智能技术来突破这些限制, 实现高逼真的数字说话人生成[1], 迅速成为了当前虚拟数字人最受关注的方向之一.

数字说话人视频 (Talking Head) 生成是当前虚拟数字人中的热点问题, 其旨在利用前沿人工智能特别是基于生成对抗网络的生成技术解决上述问题. 数字说话人视频生成可以被一般性的定义为: 给定一段驱动数据 (音频, 文本或者视频等, 如图1) 和一张包含有目标人物图像或视频, 利用这些输入中包含的信息, 合成一段包含目标人物自然表达驱动数据, 信息量通常更为丰富的视频. 其本质上是从不同模态的数据中提取信息, 在语义层面将信息扩充, 融合和对齐, 最终以直观自然的视觉形式呈现出来. 根据所提供信息类型的不同, 数字说话人生成也有着不同侧重点的子任务. 例如, 当驱动数据是文本和音频时, 生成任务侧重于生成声音, 表情和动作同步的视频. 当驱动或目标数据中进一步包含视频时, 该视频用于提取人物或场景的个性化特征, 使得所生成的视频或其中的人物特点更符合预设情境. 当前任务都以输出视频为目标, 是因为视频天然地融合了音频, 图像, 运动等多个模态的信息, 且相比其他模态更直观, 更容易被观众接受.

数字说话人视频生成技术具有广泛的应用前景, 例如电视节目播报, 电影配音, 动画制作, 线上会议, 虚拟助手等. 尤其是在虚拟数字人技术快速发展的当下, 数字说话人生成技术更是快速构建虚拟数字人不可或缺的部分. 数字说话人生成的目标是使生成的内容尽可能真实自然, 且尽可能符合驱动数据的预设情境. 相较于当前研究较多的图像到图像翻译、3D物体生成等任务, 数字说话人对生成技术提出了更高的要求, 其难点在于既需要模型从多模态数据中准确提取融合所需的语义、情感、动作等信息, 还需要将这些信息同作为参考的视觉信息有效融合, 生成自然、连续的视频.

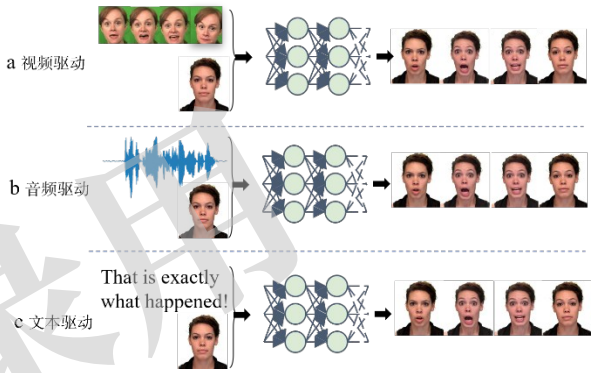


图 1 不同驱动方式的数字说话人视频生成, a 表示由视频驱动, b 表示音频驱动, c 表示文本驱动.

具体而言, 自然环境下, 人们互相交流时声音和画面是高度同步的, 这样使得人们对这种声音和画面的同步性极其敏感, 生成的视频中即使有非常微小的异常部分或者音画不匹配也很容易被观众感知. 此外, 人说话时头部动作、眼神、表情的变化都是非常丰富的, 这些内容模式都很难用现有的模型去精准地学习和控制, 所以要生成一个能够混淆人类判别、高度逼真的数字说话人视频是一个非常具有挑战性的任务. 从早期的计算机图形学结合传统机器学习的方法[2-3]到当前以深度学习为主[4-5]的方法, 该领域目前已经取得了巨大的发展, 但依然普遍存在一些问题. 例如面部细节地生成难以控制, 头部动作、表情不自然等. 此外, 随着应用场景地不断丰富, 也产生了很多需要迫切解决的新问题, 例如数据集不足, 计算速度慢, 评估策略不够有效等.

数字说话人生成的相关技术还包括了3D人脸生成[6-8]、人像风格转换[9-10]、图像生成[11-12]、语音识别[13]等任务. 针对它们的研究近年来也取得了显著的进展, 具体可查阅相关的综述工作[6-7, 11-14]. 这些任务考虑的主要是单一模态的数据, 面临的挑战是模态内建模, 需要考虑的可变因素相对有限, 解决方案也因此相对成熟. 而跨模态视频生成任务[15-16]学术界目前虽然已经开始探索, 但整体处于起步阶段. 数字说话人视频生成作为综合考虑了多种模态的跨模态生成任务, 当前自然也是处于其发展早期阶段. 虽然取得了一定进展, 但生成效果, 技术方法上仍然有很大的发展空间. 因此, 本文旨在对数字说话人视频生成当前的数据资源情况, 技术水平

状况和评估策略做一个阶段性总结, 对该领域的主要挑战和未来发展趋势进行展望. 以期能够为想要研究或者了解该领域的研究者提供一个快速的概述, 启发相关的研究和应用, 促进该领域的发展.

本文关于说话数字说话人生成的论述分为三个部分, 包括数据集, 关键技术和评估策略. 以下将逐一介绍. 最后本文将探讨数字说话人生成当前面临的挑战和未来的发展.

## 2 数据集

大规模数据是新一代人工智能能取得成功的关键因素. 作为其中的一个垂直应用领域, 数字说话人视频生成也不例外. 最初该领域只有一些基本的视频数据集, 只包含一些极短的视频及相应的字幕标注, 而且分辨率较低, 数据集的总时长较短, 包含个体也很少, 只能支持一些有限的探索, 随着越来越多的数字说话人视频数据集的公开, 数据集中包含的监督信息也越来越丰富, 如表情, 头部动作, 身体姿态视频内容等. 数据集的规模越来越大, 分辨率也越来越清晰, 这为探索该领域各种新的思路奠定了基础. 根据不同的收集方式, 数字说话人生成的数据集分别来自实验室和社交媒体两种环境下. 本文对这两类中具有代表性的数据集进行介绍, 整体情况如表 1 所示.

**表 1 数字说话人生成研究的典型数据集**

| 数据集 | 总视频时长/h | 总人数 | 有无明显头部动作 | 有无情绪标签 | 收集环境 |
|---|---|---|---|---|---|
| GRID[17] | 27.5 | 33 | — | — | 实验室 |
| CREMA-D[18] | 11.1 | 91 | — | √ | 实验室 |
| MODALITY[26] | 31.0 | 35 | — | — | 实验室 |
| MEAD[19] | 40.0 | 60 | — | √ | 实验室 |
| ObamaSet[5] | 14.0 | 1 | √ | — | 互联网 |
| LRW[20] | 173.0 | 1 k+ | √ | — | BBC |
| LRS2-BBC[21] | 224.5 | 1 k+ | √ | — | BBC |
| LRS-TED[22] | 438.0 | 5 k+ | √ | — | TED, TEDx |
| VoxCeleb[23] | 352.0 | 1 251 | √ | — | Youtube |
| VoxCeleb2[24] | 2.4 k+ | 6 112 | √ | — | Youtube |
| AVSpeech[25] | 4.9k | 150 k+ | √ | — | Youtube |

### 2.1 实验室环境下收集的数据集

GRID[17]是该领域早期一个影响力较大的语音及视频公开数据集, 由谢菲尔德大学的 Martin Cooke 等人提供. 其中包含 34 个说话个体. 每个个体分别读出 1 000 个从语料库中选出的短句, 该语料库包括 3 400 个句子. 每个句子都是从一个包含 51 个单词的集合中随机选择 6 个单词组成. 所有表演者都没有非常明显的头部动作. 在 CREMA-D[18]数据集中, 来自不同年龄段和种族的 91 位演员, 每人说出 12 个句子. 值得注意的是, 在 CREMA-D 中, 每个演员以不同类别的情绪和情绪强度来说出同一个句子, 这为探索人物表情的生成提供了数据支持. 由商汤科技等机构联合发布的 MEAD[19]据集包含 60 个演员在 8 中不同类别的情绪以及 3 种不同情绪强度下对 30 个句子表达. 此外, 该数据集还包含不同的相机视角下拍摄的图像. 与 CREMA-D 相比, MEAD 的规模更大, 总时长达 40 小时. 实验室收集的数字说话人数据集具有视频质量高, 画面清晰稳定, 音频内容清晰, 噪音干扰少. 但是其缺点也是显而易见的, 这种方式下收集的数据集受限于收集成本, 其规模都较小, 其中包含的个体较少, 内容的多样性较为单一. 此外, 由于实验室环境和真实环境差异较大, 这些数据集上训练的模型泛化性较差,应用到现实环境中通常效果不佳.

### 2.2 社交媒体环境下收集的数据集

ObamaSet[5]数据集是一个特殊的数字说话人视频数据集, 其收集了奥巴马每周的总统讲话视频. 该数据为开启特定目标说话数字人生成的技术探索提供了数据基础. 因为它只包含一个说话主体-奥巴马, 所以它无法用于训练具有泛化性的算法模型. 与之对应, LRW[20] (Lip Reading in the Wild)数据集包含数百个人读出的多达 500 个不同单词的 1000 个语句的视频. 该数据集中所有视频长度均为 29 帧 (时长约 1.16 s), 语音内容基本都出现在视频的中间. 这个数据集是从 BBC 的电视广播中收集的. LRS2-BBC[21]、LRS3-TED[22]与 LRW 类似, 分别提取来自 BBC 电视节目和 TED 和 TEDx 视频. 近年来, 互联网视频数量持续快速增加. 这为数字说话人视频生成提供了一个非常广泛的数据集来源. VoxCeleb[23]和 VoxCeleb2[24]是从 Youtube 收集的两个大规模的数字说话人数据集. VoxCeleb1 数据集包含 1 251 个个体的 100 000 多个语句视频, 而 VoxCeleb2 包含 6 112 个个体的超过一百万个语句视频. 这些从社交媒体上收集的数据集的规模普遍较大, 而且多样性较好, 同时还有多种语言, 有利于增加模型的泛化性和稳定性. 但是, 因为社交媒体的广泛性, 导致这些视频的质量也参差不齐, 这些视频普遍具有分辨率较差, 镜头抖动明显, 光照变换大和背景噪音较多等问题, 这为数据的预处理带来了很大的挑战. 此外还有一个很重要的不足: 这些视频普遍缺少必要的标签信息, 所以使用这些数据集时, 还需要进行大量的数据清洗和标注工作.

数据集的发展为数字说话人生成技术的各种探索奠定了基础. 早期的数据集由于标注信息单一, 所以该领域的早期探索较为有限, 主要关注在说话人的口型上. 随着包含更多标注信息的新数据集出现, 该领域的探索也逐步扩展到了说话人的动作, 面部表情等. 现有的数据集是在实验室和社交媒体这两种环境下收集, 在实验室环境下收集的数据集同社交媒体环境下收集的数据集在特点上存在一定的互补, 例如数据规模和标签信息等, 这种现象的存在使这两类不同的数据集可以相互弥补, 协同训练出效果更好的模型. 随着研究的不断深入, 未来对数据集规模和标注信息的需求也将不断扩增. 在未来的发展中, 研究者需要分辨率更高, 规模更大的数据集, 此外还需要带有详细的标注信息, 侧重更多细节的数据集, 例如表情, 头部动作, 眼部动作等. 此外, 数据集的结构上也将会有更多适应不同任务的设计, 例如收集成对数据适应对话式生成任务等.

## 3 关键技术

数字说话人生成虽然很早就有研究者在尝试, 但早期的方法主要是利用计算机图形学和传统机器学习的方法, 生成结果质量较差, 且存在较明显的人工合成痕迹. 随着 Alexnet[28]在图像分类中获得巨大突破, 研究者逐渐意识到深度学习对于包括数字说话人视频生成[3]内的各种视觉及多模态任务的潜力, 相关技术相比于传统方法取得了很大的进展. 当前的数字说话人生成方法普遍由三个阶段构成: (1) 特征提取与融合; (2) 特征中间表示; (3) 视频帧生成. 接下来将依据这三个阶段, 对数字说话人视频生成的主要采用的算法模型做一个梳理.

### 3.1 特征提取与融合

参考数据通常是生成目标的图像或视频, 用于提取目标人物的视觉特征, 与驱动数据提取的特征共同用于生成中间表示, 例如生成标准的 3D 人脸模型或者面部关键点. 这一小节介绍相关工作中结合数字说话人特点考虑的特征及其动机, 部分典型特征的表示方法将在 3.2 中详细叙述. 驱动数据的特征提取重点在于如何将驱动信息迁移到视觉空间中. 现在大多数方法的驱动数据都是音频数据, 少部分以文本为驱动数据的方法也都是将文本转为音频, 所以在下文的叙述中将驱动数据都默认为音频. 从信息耦合的角度来看, 音频信息中包含多种信息, 例如音频内容的信息, 发声个体的特征信息, 人物表达的情绪信息等, 每一类信息都是数字说话人生成过程中需要考虑的, 如何利用这些信息存在两种不同的观点. 一种是将整个音频看作一个整体, 只对其做简单的预处理,

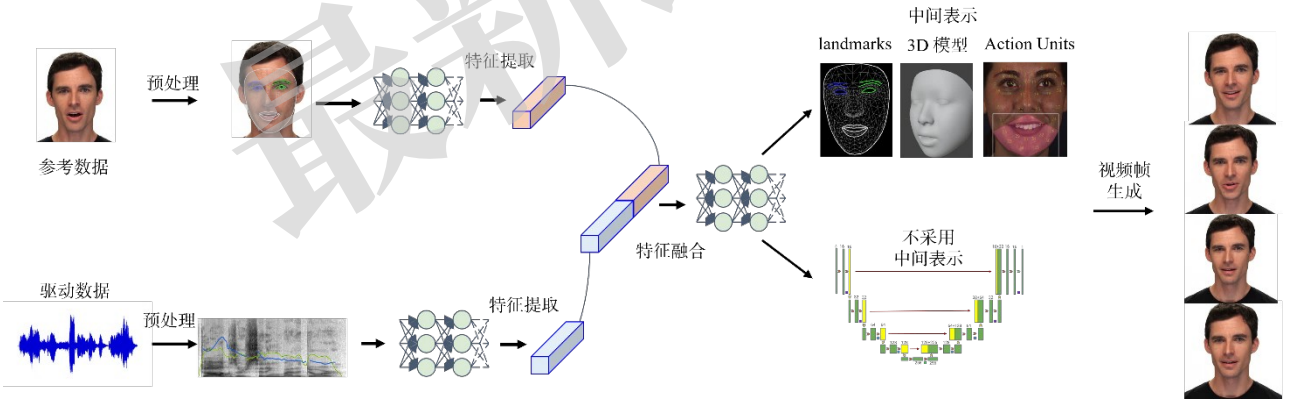


图 1 数字说话人生成方法基本流程.一般由三个环节构成: 1. 特征提取与融合, 详见第三章第一节; 2. 中间表示,详见第三章第二节; 3. 视频帧生成, 详见第三章第三节.

例如提取梅尔频谱倒谱系数(mel frequency cepstrum coefficient, MFCC,) 特征, 然后将其整个投入模型, 让模型去学习一个包含各种信息的特征. 这种方法在稍早的工作[4, 29-31]中比较常见. 文献[4]中第一次利用卷积神经网络提取音频和图像特征, 然后将不同模态的特征做融合, 接下来通过解码器对融合特征做解码, 生成视频帧. 这个方法虽然直观, 但其生成结果存在很多问题, 例如生成的视频帧模糊, 相邻帧的内容不连续, 生成的视频内容抖动明显, 保真度低等. 文献**[30]**将从音频中提取的特征同参考视频提取的标准面部关键点做融合, 生成新的面部关键点, 再利用注意力模块将新的关键点同参考图像提取的特征做融合产生新的图像和掩码图, 最后利用掩码图将新结果与参考图像融合生成最终的结果. 文献**[29]**考虑了下半部分人脸的生成, 因为嘴部区域的动作同音频信息的关联最强[30], 所以单纯只关注于这一部分的生成相较于整个人脸甚至上半身的生成更容易实现, 文献[31]比[29]增加了一个预训练的判别生成视频中音频与嘴型是否同步的专家网络(Lip-Sync Expert), 这个监督网络使得最终生成的效果有大幅提升.

另一种观点则认为音频作为语音内容、发音个体、情绪等多种信息的耦合, 如果直接将音频投入模型中, 模型很难以这种纯数据驱动的训练去学习到这些信息[32], 这样会不可避免地导致最终生成的视频存在模糊, 局部区域异常的问题. 而解决这个问题的思路是将不同类型的信息分别从音频信息中解耦出来[32-35]. 文献[32, 35]用对抗训练的方法对说话个体的特征信息和语音内容信息进行解耦, 这种思想对该领域之后的工作极具启发. 文献[33]则是直接对音频特征进行解耦, 从中解耦出内容特征和说话个体特征, 分别将这两部分特征用于关键点

的运动估计，最后将从不同特征估计的关键点运动合并起来同参考图像一起生成新的视频帧. 文献[34]则是利用 MEAD 提供的语音的情感标签，通过分解重构的方法训练模型，使得模型能够从音频信息中解耦出语音的内容特征和说话人的情绪特征.

参考数据的特征提取目标是从参考图像或视频中提取生成目标的个体信息，保证最终生成的视频人物主体是参考数据中的人物主体. 视频相比于图像数据包含更多的信息，除了必要的个体信息，还包含音频对应的头部动作，面部表情等信息，这样可以直接提取融入中间表示或者最终的视频中[35, 37]. 而参考数据是图像时，头部动作和面部表情这些动态信息是缺失的，一种思路是从音频中推理，如上文所述. 这有利于模型更好地控制最后生成内容，在控制生成细节上给了模型更多的操作空间，但信息解耦模块的设计，精细的数据标注，例如音频表达的情感类别等，更多的训练时间使得信息解耦相关的工作充满了挑战. 我们认为，关于信息解耦这一部分，未来的发展重点在两点：(1) 如何更好地解耦出包含有单一类别信息的特征；(2) 特征的可解释性. 当模型能够解析出足够好的特征时，探究每一种特征的可解释性将会对特征利用形成明确指引，从而提高生成结果的质量，同时逐步增强对于生成过程的交互性和可控性.

还有工作[36]认为从音频到头部动作，面部表情是一个多对多的映射，其中的映射关系并不明确，而且存在随机性. 所以这些工作中采用自回归[36]方式来生成头部动作和表情. 视频相比于图像能够提供更多的参考信息，但是视频的获取，预处理比图像更复杂，计算消耗也比图像更大，这也一定程度上限制了以视频为参考数据的算法的通用性和落地.

**表 2 数字说话人生成相关模型、部分技术细节及其任务范围**

| 模型 | 时间 | 整体框架 | 中间表示 | 信息解耦 | 生成个体 | 头部动作 | 面部表情 |
|---|---|---|---|---|---|---|---|
| Speech2Vid[4] | 2017 | AutoEncoder | — | — | 任意个体 | — | — |
| KIM *et al.*[38] | 2018 | AutoEncoder | 3D | √ | 任意个体 | √ | √ |
| ATVG[30] | 2019 | AutoEncoder | LandMarks | — | 任意个体 | √ | — |
| FOMM[40] | 2019 | AutoEncoder | KeyPoints | — | 任意个体 | √ | — |
| MakeItTalk[33] | 2020 | AutoEncoder | LandMarks | √ | 任意个体 | √ | — |
| Wav2lip[31] | 2020 | GAN | — | — | 任意个体 | — | — |
| FACIAL [39] | 2021 | AutoEncoder | 3D | √ | 任意个体 | √ | √ |
| LSP [36] | 2021 | AutoEncoder | FeatureMap | — | 特定个体 | √ | √ |
| AnyoneNet[41] | 2021 | AutoEncoder | Landmarks | — | 任意个体 | √ | — |
| Audio2Head [42] | 2021 | AutoEncoder | KeyPoints | — | 任意个体 | √ | — |
| EVP [34] | 2021 | AutoEncoder | 3D, Landmarks | √ | 任意个体 | √ | √ |
| PC-AVS [35] | 2021 | GAN | — | √ | 任意个体 | √ | — |
| AD-NeRF [43] | 2021 | NeRF | 3D | — | 特定个体 | √ | — |
| StyleTalker[44] | 2022 | GAN | — | — | 任意个体 | √ | √ |
| DFA-NeRF [45] | 2022 | NeRF | 3D | — | 特定个体 | √ | √ |
| DFRF[46] | 2022 | NeRF | 3D | — | 特定个体 | √ | √ |
| DiffTalk[47] | 2023 | Diffusion | — | — | 任意个体 | √ | √ |

从特征提取模型的角度来看，目前普遍采用的是卷积神经网络(convolutional neural network, CNN)和循环神经网络(recurrent neural network, RNN)，包括长短期记忆网络(long short-term memory, LSTM)[47]，门控循环单元(gate recurrent unit, GRU)[49]，其中 CNN 用于从参考图像、视频帧中提取视觉特征. 因为音频在时序上是连续的，生成的视频内容也需要这种时序的连续性，所以 RNN 也普遍用于从语音这些时序的驱动数据中提取特征，以期从中学习到这种时序的连续性，进而保证生成内容在时序上的连续性. 当前数字说话人视频生成方法中设计的子模型一般都较为简单，原因是数字说话人生成本质上是一个视频生成任务，如果模型的结构太复杂，计算的效率会大大降低. 同样，这种计算效率的问题也限制了其他视觉任务中的新结构推广到数字说话人生成任务中来，这也是制约该领域发展的一个非常重要的因素. 要解决这个问题，一方面依赖于硬件性能的提高，另一方面则是模型的改进例如模型的轻量化. 未来，能够实时生成的数字说话人方法[28]将会是一个研究者重点关注的方向.

目前说话数字人生成技术中，对于不同模态特征的融合都是采用简单的线性运算或者维度拼接，然后通过一个网络来隐式的学习不同模态特征的语义融合，最后将融合后的特征映射到中间表示或者

直接加入视频帧生成的子模型中. 总体来说目前的融合方式比较单一, 融合的机制也较为简单, 恐不能充分挖掘利用不同模态之间的相关性, 所以探索新的特征融合机制也是未来该领域一个需要考虑的问题.

### 3.2 特征中间表示

数字说话人视频生成的方法大多都会采取一定的中间表示来辅助生成, 中间表示一般都是利用现有的算法通过参考数据生成一个目标人物的人脸模型, 将从驱动数据中学习到的特征映射到这个模型上, 然后再通过驱动信息修改后的人脸模型来生成说话数字人视频. 相比于不采用中间表示的方法[4, 32], 中间表示能够给最终的生成阶段提供人脸的结构信息, 约束生成过程, 保证生成内容的质量和可控性. 数字说话人生成中用到的中间表示有二维面部关键点(landmarks)[30, 33-34, 77]和3D人脸模型[5, 39, 50-51], 其次还有面部动作单元[61](action unit, AU)标签等.

二维面部关键点是对人脸特定位置进行标注的点集. 常见的面部特征点包括对人脸轮廓、五官等的标注. 面部关键点的获取由面部关键点检测器来获取. 面部关键点检测算法比较成熟, 有很多方便的算法库[52], 例如dlib[53]. 所以二维的面部特征点是一个非常方便的面部表示. 面部关键点包含的数据比较多时, 可以采用降维的方式, 将标准的面部关键点数据降维到一个较低的维度, 同驱动数据中提取的音频特征做融合后恢复到原先的维度, 获取新的面部特征点[30]. 文献[32]中采用预测每个特征点的位移来代替直接预测特征点的位置, 这样有利于提高预测的准确度. 此外, 有些工作也用热力图来表示每个特征点[40].

另一个常用的中间表示是3D人脸模型. 数字说话人视频生成中基本都是采用参考图像重构3D模型, 重构算法大多都是采用3D可变形人脸模型(3D morphable models, 3DMM)[54]. 该算法的核心思想是每个人脸的3D模型都可以在3D空间中进行匹配, 由其他人脸的正交基线性加权获得. 文献[54]中提出的3DDFA (3D dense head alignment, 3DDFA)算法将人脸模型表示为姿态, 形状和表情的加权和. $S$ 表示人脸模型, $E_s$, $E_e$ 表示形状和表情的基向量, $W_s$, $W_e$ 表示对应的系数. 每个人脸模型表示为:

$$S = \bar{S} + E_s W_s + E_e W_e$$

对于数字说话人来说, 3DMM算法的优势在于分解了人脸的形状、表情等属性, 这很大程度上简化了对于头部动作、表情的操作, 所以目前大多数数字说话人生成的方法都是直接用从驱动数据中提取的特征计算3DDFA输出的某一项参数, 例如表情, 然后用这些参数结合基础的3D人脸重新计算一个新的3D人脸模型[39, 50-51]. 但是当前3DMM的相关模型受限于高精度数据集缺少导致的对于面部细节的建模效果并不是很理想, 所以采用3DMM方法的数字说话人生成方法都有一些辅助的模块来对面部细节, 例如眼角、嘴角等重新进行细化. 神经辐射场(neural radiance fields, Nerf)[56]是另一个较新的3D模型的建模方案. 其思想是通过一个神经网络来学习某一物体不同视角的2D图像到3D模型的映射. 假设某个场景中某一个空间点的坐标 $p = (x, y, z)$, 视角方向为 $d = (\theta, \phi)$, 用于学习神经网络的映射如下:

$$F_\theta = (p, d) = (c, \sigma)$$

其中 $c = (r, g, b)$ 表示从该视角观察到该点的颜色, $\sigma$ 则是该点的密度. 根据传统的体渲染方法, 从3D场景的渲染出的2D图像中每个像素对应一条从该像素发出的光线上所有空间点的颜色和密度的积分得到. 而Nerf则在渲染过程中用神经网络预测该射线上每个点的颜色和密度. 受这种方式的启发, 文献[43, 45-46]也采用了这种方式实现数字人的建模. 但这种方式有两个明显的不足, 第一个是每个Nerf模型对应一个具体的数字人个体, 无法实现跨个体的数字人建模; 第二个是渲染速度较慢. 原始的Nerf在渲染一帧需要一分钟左右, 随着相关技术的不断发展进步, 相信这两方面的不足在未来一段时间内也将得到有效解决.

面部动作编码系统(Facial Actin Coding System, FACS)[57]是由Paul Ekman和Wallace V. Friesen在前人基础上开发的用于描述人脸表情的一个标签系统, 其在2002年做了大范围扩充. 该系统根据解剖学将人脸分为若干个区域, 每个区域的每个状态都有一个固定编码, 每一种人脸表情都可以有一个或多个AU编码来表示. 有一部分侧重人脸表情的数字说话人生成的工作[61]采用这套系统作为人脸的部分中间表示. 除了作中间表示, 这套系统还可以作为标注信息用于定义人脸表情.

总而言之, 中间表示作为对于生成过程的信息扩充, 可以采用一种中间表示, 也可采用多种来共同表征人脸, 甚至在保证生成结果的前提下可以不采用中间表示. 之后更多的侧重点应该是关注不同中间表示优点的融合, 在保证结果的前提下尽可能地减少计算消耗.

### 3.3 视频帧生成

视频帧生成的目标是利用从输入数据中提取的特征生成数字人视频, 本质上是一个跨模态的视频生成任务. 一个直观的解决方案是使用一个模型从驱动数据直接生成视频, 但实际发展中发现这种简单的解决方案存在许多问题. 其中最重要的问题是缺乏人脸结构信息的约束, 容易出现生成的图像失真或主体特征变化等. 对此, 主流的解决方案是引入3.2小节中介绍的中间表示, 通过它们来约束生成过程. 但引入中间表示也带来了一些其他的问题, 例如中间表示的精确度问题导致的信息损失, 人脸属性的解耦. 另一种解决方案是利用预训练模型来实现视频生成, 例如在大规模人脸数据集上预训练的StyleGAN[58]作为生成器来生成视频帧, 此外这种方式还可以通过隐向量的变化来实现对生成内容的控制, 但这种方法的缺点在于大模型本身是

黑盒模型, 整个生成过程都无法显式地进行控制, 如果变化幅度过大很容易出现内容畸变等异常. 在具体的实现上, 如果使用 3D 模型作为表示, 则可以通过渲染引擎直接渲染出视频帧. 此外, 也可以利用生成对抗网络(generative adversarial network, GAN)[59]或其他卷积神经网络来生成视频帧. 最近, 扩散模型 (Diffusion Model)[60]在生成领域的发展迅速, 该模型基于随机漫步和梯度反演的思想来实现生成. 首先, 扩散模型建立一个随时间演化的概率分布, 该分布在每个时间步骤中会被加入一些噪声, 以增加样本的多样性. 通过对该概率分布进行迭代采样, 可以生成与训练数据具有相似统计特征的新样本. 相比 GAN, 其优化过程更清晰, 训练更稳定, 生成结果的多样性也更好, 尤其是在多模态生成领域的成功对数字人生成非常有参考价值. 总体来说, 现有方法的生成结果在内容的真实性, 流畅度等方面还有许多不足, 而现有很多生成方法因为数据, 算力等方面的限制难以融入数字说话人视频的生成中, 所以我们认为如何将生成领域的最新方法合理到应用到数字人视频生成是未来该方向一个重要的研究内容.

## 4 评估策略

### 4.1 主观评估

主观评估是对生成任务最直观有效的评估方式. 在数字说话人生成中, 主观评估的作用尤为重要. 对于一些生成的异常情况, 例如头部动作畸形, 面部表情僵硬, 数字说话人动作同语音内容、情感不匹配等, 目前都没有很好的客观指标来评估, 所以在说话数字人生成任务中主观评估依然有着不可替代的作用. 评估的方式一般是将不同来源的视频, 例如不同方法生成的视频以及真实视频, 给评估者随机观看, 然后由评估者对这些视频进行打分, 然后根据最终的结果计算相关的统计指标, 来评估模型的生成结果的优劣. 在评价过程的设计中, 平均主观意见分 (Mean Opinion Score, MOS) 是一个常用的方法, 即对大量评估者的评分取平均值来表示最终的结果. 虽然主观评估是对生成结果最真实的反映, 但是主观评估有一个非常明显的缺点, 就是每个参与评价的评估者的感知, 认同等都有差异, 而这种差异不可避免地会体现在统计结果中, 造成偏差, 为了消除这种偏差, 只能请更多的评估者进行评估. 而这样需要的评估成本是很难承担的, 特别是在需要对比多个方法的情况. 因此, 更加全面的客观评估评估方案是目前该领域一个非常迫切需要解决的问题.

### 4.2 客观评估

目前的客观评估策略注重于评估音频和视频的一致性以及生成视频帧的质量. 评估音视频一致性的指标目前侧重于评估生成的嘴型和音频是否对应, 评估生成质量的指标沿用了图像生成任务的指标. 下面对这两方面中应用较多的几个指标做详细阐述.

1. 峰值信噪比

峰值信噪比(peak signal to noise radio, PSNR)[65]通常用于图像压缩及其他信号领域的信号重构质量评估中, 在数字说话人生成中用于评估生成视频帧的质量. PSNR 通过均方差(mean square error, MSE)来定义, 给定一张大小为 $m\times n$ 单通道图像 $\boldsymbol{I}$ 和一张带有噪音的近似图像 $\boldsymbol{K}$, PSNR 定义如下:

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

$$\mathrm{PSNR}(\boldsymbol{I},\boldsymbol{K})=10\lg\left(\frac{\mathrm{MAX}(\boldsymbol{I})}{\mathrm{MSE}(\boldsymbol{I},\boldsymbol{K})}\right)=20\lg\left(\frac{\mathrm{MAX}^2(\boldsymbol{I})}{\sqrt{\mathrm{MSE}(\boldsymbol{I},\boldsymbol{K})}}\right)$$

PSNR 在评估图像质量中应用非常广泛, 但局限性也很明显, 由于 PSNR 根据 MSE 计算, 当图像内容不同时, 图像的 MSE 可能相同, 即 PSNR 值相同的图像在人眼中可能差异明显. 换句话说, PSNR 和人眼的感知效果并不完全一致[52].

2. 结构相似性

结构相似性(structural similarity index metric, SSIM)[66]从亮度, 对比度和结构三个方面来评估图像的相似度. 给定两幅图像 $\boldsymbol{I}$ 和 $\boldsymbol{E}$, 其均值分别定义为 $\mu_I$ 和 $\mu_E$, 标准差为 $\sigma_I$ 和 $\sigma_E$, 协方差定义为 $\sigma_{IE}$. 基于此, 其亮度 $l$、对比度 $c$、结构的相似度 $s$ 定义分别如下式.

$$l(\boldsymbol{I},\boldsymbol{E})=\frac{2\mu_I\mu_E+c_1}{\mu_I^2+\mu_E^2+c_1}$$

$$c(\boldsymbol{I},\boldsymbol{E})=\frac{2\mu_I\mu_E+c_2}{\sigma_I^2+\sigma_E^2+c_2}$$

$$s(\boldsymbol{I},\boldsymbol{E})=\frac{\sigma_{IE}+c_3}{\sigma_I\sigma_E+c_3}$$

其中 $c_1$, $c_2$, $c_3$ 是常量. 分别如下式:

$$c_1=(k_1L)^2,\ c_2=(k_2L)^2,\ c_3=c_2/2$$

$k_1=0.01$, $k_2=0.03$, $L$ 是图像中可能出现的最大值, 例如, 图像中每个像素值为 8bit, 则该图像的 $L$ 值为 255. 最后, 结构相似性定义如下:

$$\mathrm{SSIM}(\boldsymbol{I},\boldsymbol{E})=l(\boldsymbol{I},\boldsymbol{E})c(\boldsymbol{I},\boldsymbol{E})s(\boldsymbol{I},\boldsymbol{E})$$

相比于 PSNR, SSIM 的评测效果更接近人类视觉观察的结果. SSIM 应用在局部图像质量的评估上效果更好, 这也更符合人眼在观测图像时每次都只能关注局部区域的特点. SSIM 除了作为评估策略外, 其还常见于构造损失函数. SSIM 损失如下:

$$L_{\mathrm{SSIM}}=1-\mathrm{SSIM}(\boldsymbol{I},\boldsymbol{E})$$

3. Frechet Inception Distance

FID(frechet inception distance, FID)[27]用于评

估生成模型生成的图像质量. FID 通过比较真实图像和生成图像的分布差异来计算. FID 表示的是两个多维高斯分布的 Wasserstein 距离的平方根. FID 需要一个预训练的神经网络对参与计算的图像做特征提取, 一般都采用在 ImageNet 数据集上预训练的 Inceptionv3[67]网络作为特征提取网络. 定义真实图像 $I$ 的特征分布为 $N(\mu, \Sigma)$, 生成图像 $\hat{I}$ 的特征分布为 $N(\hat{\mu}, \hat{\Sigma})$. FID 定义如下:

$$\mathrm{FID}(I, \hat{I}) = \|\mu - \hat{\mu}\|_2^2 + \mathrm{tr}\left(\Sigma + \hat{\Sigma} - 2(\Sigma\hat{\Sigma})^{\frac{1}{2}}\right)$$

FID 在卷积网络提取特征上进行计算, 更侧重于图像的语义信息. 但也有其局限性, 首先 FID 基于真实数据和生成数据的特征分布差异, 因此对真实数据分布的假设非常敏感, 其次无法捕捉细节和多样性, FID 指标关注整体的相似性, 而忽视了生成样本的细节和多样性, 而数字人视频生成任务对细节的要求又非常高.

4. 关键点距离

关键点距离 (landmark distance)[69]是一个简单但有效的评估指标, 其利用真实视频和生成视频的嘴部关键点的欧氏距离来衡量生成嘴型的准确性, 计算方式如下:

$$\mathrm{LMD}(v, \hat{v}) = \frac{1}{TP}\sum_{t=1}^{T}\sum_{p=1}^{P}\left\|\mathrm{Dis}(v) - \mathrm{Dis}(\hat{v})\right\|_2^2$$

其中 $T$ 表示视频长度, $P$ 表示关键点数量, $v$ 和 $\hat{v}$ 表示真实视频和生成视频, $\mathrm{Dis}$ 表示关键点检测函数. 这个指标除了应用于嘴型之外也可扩展至整个面部来评估面部表情. 但是如果生成的内容包括头部动作时, 需要通过对齐来移除头部动作的影响.

5. SyncNet 评分

SyncNet 评分 (SyncNet Score)[70]是通过对比由神经网络提取的音频和视频特征来评估视频和音频的一致性. 计算时通过两个预训练网络分别提取音频特征和视频特征, 然后计算两类特征的距离, 之后再根据距离来计算音频和视频的时间差 (offset) 以及置信度 (confidence), 具体的计算如下:

$$d = \|f_v - f_a\|_2^2$$

$$o = I - \arg\min(d)$$

$$s_{conf} = \mathrm{Md}(d) - \min(d)$$

其中 $o$ 表示 offset, $s_{conf}$ 表示 confidence, $f_v$ 和 $f_a$ 分别为神经网络提取的视频特征和音频特征, $I$ 表示常数, $\mathrm{Md}$ 表示中位数.

总的来说, 除主观评估外, 现有的数字说话人视频生成方法采取的客观评估指标侧重于图像质量和嘴型同音频一致性, 除了上述这几种方法之外, 还存在一些别的测评指标. 但是对于视频的内容的连续性、合理性等, 目前并没有广泛认可的客观指标来进行衡量. 尽管新的方法被不断提出[68], 但是当前依然缺乏一个被大多数研究者认可的比较全面有效的客观评估策略.

## 5 挑 战

前面从数据集, 关键技术, 评估策略三个方面对数字说话人视频生成的当前状况做了系统的总结. 总的来说, 当前生成效果距离真实视频依然有不小的差距. 此外, 在数据集, 计算效率, 评估指标等也还存在一些需要解决的挑战. 总体来说, 当前该领域面临的挑战有以下几点:

1. 提升生成视频的细节质量. 数字说话人生成的结果是视频, 影响视频质量的因素较多. 现存的算法仅兼顾到了一些主要的特征, 例如嘴型、头部动作等. 这样造成了生成结果的表现力不足, 人物动作、表情、眼神等呆板, 不生动. 要改善这一点则需要模型更侧重于细节的生成, 尤其是一些随机性较强的细节, 例如眨眼, 面部的微表情等. 这一方面依赖新的方法[71-72]的发展. 另一方面, 这些细节需要大量的训练数据. 我们认为, 利用迁移学习的方法, 将在大规模数据集上通过自监督学习预训练得到的大模型迁移过来[73], 再根据任务需求, 如利用公开的面部微表情数据[74-75]等, 对大模型进行微调是解决这个问题值得探索的方向之一.

2. 构建更高质量的数据集. 现有的公开数据集无论是在质量还是数据规模上都存在一定不足. 实验室收集的高质量数据集由于成本的原因限制了其规模, 社交媒体上收集的数据集虽然规模较大, 但质量参差不齐, 同时还缺乏细粒度的标注. 未来数字说话人生成数据集除了要扩大数据集中的个体数目, 增加整体的视频时长, 还应更侧重于高质量, 细粒度的数据收集和标注. 此外, 借鉴一些相关领域的技术方法来对现有的数据集进行优化, 例如用现有的表情识别网络对数据进行初步标注[76], 用生成模型对数据集的某个固定特征进行迁移, 获得满足任务要求的数据集[77]. 充分利用人机融合的方式, 在较小人力投入的情况下实现数据的高质量标注.

3. 降低生成模型的计算复杂度. 视频生成任务的高计算代价导致许多前沿的视觉计算模型, 例如 Transformer[78]、StyleGAN[58]等, 难以有效推广到该领域, 阻碍了该领域的发展. 所以, 降低生成过程的资源消耗是当前该领域需要迫切解决的问题之一. 例如, 设计轻量化的模型或是对模型进行轻量化. 除了对模型的结构和参数进行优化外, 还可以用一些其他的策略, 例如知识蒸馏[79], 模型剪枝[80]来减少模型的计算消耗. 同时, 结合视频的时空连续性, 冗余度高的特点设计可有效降低数据处理量的方法也是一个值得关注的方向.

4. 发展更全面的评估指标. 当前的评估策略侧重在图像层次关注生成视频的质量, 对于视频内容整体的自然度, 动作连贯性, 音频与视频的一致性, 合理性等缺乏有效的客观评估手段, 例如现有的指标无法反映出生成的说话人面部表情只有少数固定的模式或者头部动作循环这种异常情况, 但对于真

人, 这种缺陷很容易观察到. 评估手段的不足极大阻碍了数字说话人技术的发展. 我们认为需要发展新的评估指标, 可以对生成内容的上述几个维度进行有效评估, 解决当前客观评估方法都难以有效评估的问题.

## 6 总 结

本文从数据集, 关键技术和评估策略三个方面, 对当前数字说话人生成技术做了一个阶段性总结. 以外, 还梳理了当前数字说话人生成面临的主要挑战, 提出了我们对于如何突破这些瓶颈的见解. 希望这篇文章能够对想要进入该领域或者研究该领域的学者提供一个快速的了解和启发, 促进该领域的发展.

数字说话人生成技术已经获得了阶段性的发展, 以深度学习为代表的方法技术快速迭代更新也使得生成的视频越来越逼真. 在数字媒体中, 数字说话人也逐步进入应用阶段, 很多视频平台陆续开始尝试部署数字说话人生成的功能, 虚拟主持、虚拟助手也不断出现在日常的生活中. 当前, AIGC (AI-Generated Content) 技术也取得了长足进步, 尤其是在跨模态图像生成[15, 81-82], 文本生成[83]等领域产生巨大影响. 虽然跨模态的视频生成任务目前还有待探索, 但随着新方法的不断涌现, 多模态的视频生成任务也将成为 AIGC 的一个重点探索的任务. 随着新的方法的出现和计算机算力的提升, 数字说话人生成也将迎来一个新的发展阶段. 未来, 数字说话人可能突破拘泥于现阶段的任务要求, 仅需要更少的信息, 例如一段描述性的文本即可生成视频. 随着阻碍该技术发展的瓶颈不断被新的方法突破. 有理由相信, 数字说话人生成技术除了广泛存在于虚拟世界外, 还会同现实中的机器人等深度融合, 为用户提供更加贴近真实的交互体, 创造出更多具有创造力和价值的应用.

## 参考文献(References):


[1] Korban M, Li X. A Survey on Applications of digital human avatars toward virtual co-presence[J]. arXiv preprint arXiv: 2201.04168, 2022.

[2] Wang L, Qian X, Han W, *et al.* Synthesizing photo-real talking head via trajectory-guided sample selection[C]//Eleventh Annual Conference of the International Speech Communication Association. 2010.

[3] Xie L, Liu Z Q. Realistic mouth-synching for speech-driven talking face using articulatory modelling[J]. IEEE Transactions on Multimedia, 2007, 9(3): 500-510.

[4] Chung J S, Jamaludin A, Zisserman A. You said that?[J]. arXiv preprint arXiv:1705.02966, 2017.

[5] Suwajanakorn S, Seitz S M, Kemelmacher-Shlizerman I. Synthesizing obama: learning lip sync from audio[J]. ACM Transactions on Graphics, 2017, 36(4): 1-13.

[6] Sharma S, Kumar V. 3D face reconstruction in deep learning era: A Survey[J]. Archives of Computational Methods in Engineering, 2022, 29(5): 3475-3507.

[7] Gao T, An H. A Brief Survey: 3D Face Reconstruction [C]//International Conference on Broadband and Wireless Computing, Communication and Applications. Springer, Cham, 2019: 846-854

[8] 沈铖潇, 钱丽萍, 俞宁宁. 融合 UV 位置图与 CGAN 的单图像大视角三维彩色人脸重建 [J]. 计算机辅助设计与图形学学报, 2022, 34 (04) : 614-622.
Shen Chengxiao, Qian Liping, Yu Ningning. Large-View 3D Color Face Reconstruction from Dingle Image via UV Location Map and CGAN[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(4): 614-622. doi: 10.3724/SP.J.1089.2022.18959

[9] Amodio M, Krishnaswamy S. Travelgan: Image-to-image translation by transformation vector learning [C]//Proceedings of the ieee/cvf conference on computer vision and pattern recognition. 2019: 8983-8992

[10] Wang C, Xu C, Wang C, *et al.* Perceptual adversarial networks for image-to-image transformation[J]. IEEE Transactions on Image Processing, 2018, 27 (8): 4066-4079

[11] Wu X, Xu K, Hall P. A survey of image synthesis and editing with generative adversarial networks[J]. Tsinghua Science and Technology, 2017, 22(6): 660-674.

[12] Xia W, Zhang Y, Yang Y, *et al.* Gan inversion: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(3): 3121-3138

[13] Malik M, Malik M K, Mehmood K, *et al.* Automatic speech recognition: a survey[J]. Multimedia Tools and Applications, 2021, 80: 9411-9457.

[14] Chen Y, Zhao Y, Jia W, *et al.* Adversarial-learning-based image-to-image transformation: A survey[J]. Neurocomputing, 2020, 411: 468-486.

[15] Ramesh A, Dhariwal P, Nichol A, *et al.* Hierarchical text-conditional image generation with clip latents[J]. arXiv preprint arXiv:2204.06125, 2022.

[16] Hong W, Ding M, Zheng W, *et al.* Cogvideo: Large-scale pretraining for text-to-video generation via transformers[J]. arXiv preprint arXiv:2205.15868, 2022.

[17] 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): 2421-2424.

[18] Cao H, 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.

[19] Wang K, Wu Q, Song L, *et al.* Mead: A large-scale audio-visual dataset for emotional talking-face generation[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI. Cham: Springer International Publishing, 2020: 700-717.

[20] Chung J S, Zisserman A. Lip reading in the wild[C]//Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13. Springer International Publishing, 2017: 87-103.

[21] Chung J S, Senior A, Vinyals O, *et al.* Lip reading sentences in the wild[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 6447-6456.

[22] Afouras T, Chung J S, Zisserman A. LRS3-TED: a large-scale dataset for visual speech recognition[J]. arXiv preprint arXiv:1809.00496, 2018.

[23] Nagrani A, Chung J. S., Zisserman A. Voxceleb: a large-scale speaker identification dataset[J]. In Interspeech, 2017: 2616-2620.

[24] Chung, J S, Nagrani, A, Zisserman, A (2018) VoxCeleb2: Deep Speaker Recognition[J]. In Interspeech 2018: 1086-1090.

[25] Ephrat A, Mosseri I, Lang O, *et al.* Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation[J]. ACM Transactions on Graphics 37(4): 112:1-112:11.

[26] Czyzewski A, Kostek B, Bratoszewski P, *et al.* An audio-visual corpus for multimodal automatic speech recognition[J]. Journal of Intelligent Information Systems, 2017, 49: 167-192.
[27] Heusel M, Ramsauer H, Unterthiner T, *et al.* Gans trained by a two time-scale update rule converge to a local nash equilibrium[J]. Advances in neural information processing systems, 2017, 30.
[28] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.
[29] KR P, Mukhopadhyay R, Philip J, *et al.* Towards automatic face-to-face translation[C]//Proceedings of the 27th ACM international conference on multimedia. 2019: 1428-1436.
[30] Chen L, Maddox R K, Duan Z, *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. 2019: 7832-7841.
[31] 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. 2020: 484-492.
[32] Zhou H, Liu Y, Liu Z, *et al.* Talking face generation by adversarially disentangled audio-visual representation[C]//Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 9299-9306.
[33] Zhou Y, Han X, Shechtman E, *et al.* Makelttalk: speaker-aware talking-head animation[J]. ACM Transactions On Graphics , 2020, 39(6): 1-15.
[34] Ji X, Zhou H, Wang K, *et al.* Audio-driven emotional video portraits[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 14080-14089.
[35] Zhou H, Sun Y, 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. 2021: 4176-4186.
[36] Lu Y, Chai J, Cao X. Live speech portraits: real-time photorealistic talking-head animation[J]. ACM Transactions on Graphics, 2021, 40(6): 1-17.
[37] Lahiri A, Kwatra V, Frueh C, *et al.* Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 2755-2764.
[38] Kim H, Garrido P, Tewari A, *et al.* Deep video portraits[J]. ACM Transactions on Graphics, 2018, 37(4): 1-14.
[39] Zhang C, Zhao Y, Huang Y, *et al.* Facial: Synthesizing dynamic talking face with implicit attribute learning[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 3867-3876.
[40] Siarohin A, Lathuilière S, Tulyakov S, *et al.* First order motion model for image animation[J]. Advances in Neural Information Processing Systems. 2019: 7137–7147.
[41] Wang X, Xie Q, Zhu J, *et al.* AnyoneNet: Synchronized Speech and Talking Head Generation for Arbitrary Persons[J]. IEEE Transactions on Multimedia, 2022.
[42] Wang S, Li L, Ding Y, *et al.* Audio2head: Audio-driven one-shot talking-head generation with natural head motion[J]. arXiv preprint arXiv:2107.09293, 2021.
[43] Guo Y, Chen K, 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. 2021: 5784-5794.
[44] Min D, Song M, Hwang S J. StyleTalker: One-shot Style-based Audio-driven Talking Head Video Generation[J]. arXiv preprint arXiv:2208.10922, 2022.
[45] Yao S, Zhong R Z, Yan Y, *et al.* DFA-NERF: personalized talking head generation via disentangled face attributes neural rendering[J]. arXiv preprint arXiv:2201.00791, 2022.
[46] Shen S, Li W, Zhu Z, *et al.* Learning dynamic facial radiance fields for few-shot talking head synthesis [C]//Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII. Cham: Springer Nature Switzerland, 2022: 666-682.
[47] Shen S, Zhao W, Meng Z, *et al.* DiffTalk: Crafting Diffusion Models for Generalized Talking Head Synthesis[J]. arXiv preprint arXiv:2301.03786, 2023.
[48] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
[49] Cho K, Van Merriënboer B, Gulcehre C, *et al.* Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.
[50] Thies J, Elgharib M, Tewari A, *et al.* Neural voice puppetry: Audio-driven facial reenactment[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16. Springer International Publishing, 2020: 716-731.
[51] Chen L, Cui G, Liu C, *et al.* Talking-head generation with rhythmic head motion[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX. Cham: Springer International Publishing, 2020: 35-51.
[52] Khabarlak K, Koriashkina L. Fast facial landmark detection and applications: A survey[J]. arXiv preprint arXiv:2101.10808, 2021.
[53] King D E. Dlib-ml: A machine learning toolkit[J]. The Journal of Machine Learning Research, 2009, 10: 1755-1758.
[54] Blanz V, Vetter T. A morphable model for the synthesis of 3D faces[C]//Proceedings of the 26th annual conference on Computer graphics and interactive techniques. 1999: 187-194.
[55] Zhu X, Lei Z, Liu X, *et al.* Face alignment across large poses: A 3d solution[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 146-155.
[56] Mildenhall B, Srinivasan P P, Tancik M, *et al.* Nerf: Representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2021, 65(1): 99-106.
[57] Ekman P, Friesen W V. Facial action coding system[J]. Environmental Psychology & Nonverbal Behavior, 1978.
[58] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4401-4410.
[59] Goodfellow I, Pouget-Abadie J, *et al.* Generative adversarial nets[J]. Advances in neural information processing systems. 2014: 2672–2680.
[60] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.
[61] Chen S, Liu Z, Liu J, *et al.* Talking head generation with audio and speech related facial action units[J]. arXiv preprint arXiv:2110.09951, 2021.
[62] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015: 234-241.
[63] Isola P, Zhu J Y, Zhou T, *et al.* Image-to-image translation with conditional adversarial networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1125-1134.
[64] Kato H, Beker D, Morariu M, *et al.* Differentiable rendering: A survey[J]. arXiv preprint arXiv:2006.12057, 2020.

[65] Huynh-Thu Q, Ghanbari M. Scope of validity of PSNR in image/video quality assessment[J]. Electronics letters, 2008, 44(13): 800-801.
[66] 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.
[67] Szegedy C, Liu W, Jia Y, *et al.* Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.
[68] Chen L, Cui G, Kou Z, *et al.* What comprises a good talking-head video generation?[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020.
[69] Chen L, Li Z, Maddox R K, *et al.* Lip movements generation at a glance[C]//Proceedings of the European conference on computer vision. 2018: 520-535.
[70] Chung J S, Zisserman A. Out of time: automated lip sync in the wild[C]//Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13. Springer International Publishing, 2017: 251-263.
[71] Dosovitskiy A, Beyer L, Kolesnikov A, *et al.* An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.
[72] Liu Z, Lin Y, Cao Y, *et al.* Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.
[73] Tan C, Sun F, Kong T, *et al.* A survey on deep transfer learning[C]//Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27. Springer International Publishing, 2018: 270-279.
[74] Yan W J, Wu Q, Liu Y J, *et al.* CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces[C]//2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG). IEEE, 2013: 1-7.
[75] Ben X, Ren Y, Zhang J, *et al.* Video-based facial micro-expression analysis: A survey of datasets, features and algorithms[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 44(9): 5826-5846.
[76] Devi B, Preetha M M S J. A Descriptive Survey on Face Emotion Recognition Techniques[J]. International Journal of Image and Graphics, 2021: 2350008.
[77] Tzaban R, Mokady R, Gal R, *et al.* Stitch it in time: Gan-based facial editing of real videos[C]//SIGGRAPH Asia 2022 Conference Papers. 2022: 1-9.
[78] Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
[79] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network [J]. arXiv preprint arXiv: 1503. 02531, 2015, 2 (7)
[80] Li H, Kadav A, Durdanovic I, *et al.* Pruning filters for efficient convnets[J]. arXiv preprint arXiv:1608.08710, 2016.
[81] Radford A, Kim J W, Hallacy C, *et al.* Learning transferable visual models from natural language supervision [C]//International conference on machine learning. PMLR, 2021: 8748-8763.
[82] 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. 2022: 10684-10695.
[83] Ouyang L, Wu J, Jiang X, *et al.* Training language models to follow instructions with human feedback[J]. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.