计算机科学
COMPUTER SCIENCE

说话人生成研究现状与发展趋势

宋昕洋, 阎志远, 孙沐毅, 戴琳琳, 李琦, 孙哲南

引用本文

宋昕洋, 阎志远, 孙沐毅, 戴琳琳, 李琦, 孙哲南说话人生成研究现状与发展趋势[J]. 计算机科学, 2023, 50(8): 68-78.

SONG Xinyang, YAN Zhiyuan, SUN Muyi, DAI Linlin, LI Qi, SUN Zhenan. Review of Talking Face Generation [J]. Computer Science, 2023, 50(8): 68-78.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221000031

说话人生成研究现状与发展趋势

宋昕洋[1,2]　阎志远[3]　孙沐毅[2]　戴琳琳[3]　李　琦[1,2]　孙哲南[1,2]
1 中国科学院大学人工智能学院　北京 100049
2 中国科学院自动化研究所模式识别国家重点实验室智能感知与计算研究中心　北京 100190
3 中国铁道科学研究院集团有限公司电子计算技术研究所　北京 100081
(songxinyang2022@ia.ac.cn)

摘　要　说话人生成是视觉生成领域的热门研究方向,旨在根据输入的多模态信息生成逼真的说话人视频。说话人生成在影视传媒、游戏动漫和互联网相关产业中具有广阔的应用前景,同时也可以为唇读识别、伪造鉴别和数字人生成等任务的研究提供数据支持。现阶段主流的说话人生成方法已经能够实现包含个性化属性、视听同步的说话人视频生成,但还未能达到虚拟现实、人机交互和元宇宙等新兴应用场景的要求。因此,研究说话人生成对于推动相关产业发展具有重要意义。对说话人生成的研究现状进行梳理与总结,首先阐述了说话人生成的研究背景和相关技术,然后根据方法分类介绍了近年来主流的说话人生成方法,整理了相关研究中常用的视听数据集和评价指标,最后总结现有方法存在的问题,分析了说话人生成未来潜在的研究方向。

关键词:人脸生成;视频生成;图像生成;深度学习;多模态学习;人脸重建;深度伪造;计算机视觉
中图法分类号　TP391

Review of Talking Face Generation

SONG Xinyang[1,2],YAN Zhiyuan[3],SUN Muyi[2],DAI Linlin[3],LI Qi[1,2] and SUN Zhenan[1,2]
1 School of Artificial Intelligence,University of Chinese Academy of Sciences,Beijing 100049,China
2 Center for Research on Intelligent Perception and Computing,National Laboratory of Pattern Recognition,Institute of Automation,Chinese Academy of Sciences,Beijing 100190,China
3 Institute of Computing Technology,China Academy of Railway Sciences Corporation Limited,Beijing 100081,China

Abstract　Talking face generation is a popular research direction in the field of visual generation,which aims to generate realistic speaker videos based on multimodal input data. Talking face generation has broad application prospects in video media,game animation and Internet-related industries,and it could also provide data support for the research of tasks such as lip reading recognition,fake identification and digital human generation. The existing mainstream methods have been able to achieve talking face generation with personalized attributes and audio-visual synchronization,but they fail to meet the requirements of emerging application scenarios such as virtual reality,man-machine interaction and metaverse. Sothestudyof talking face generation is of great significance for promoting the development of related industries. This paper sorts out and summarizes the research status of talking face generation. First,itelaborates the research background and related technologies of talking face generation,then introduces the mainstream generation methods in recent years according to the method classification,sorts out the audio-visual datasets and evaluations commonly used in the research,and finally summarizes the problems in the existing methods,and analyzes the potential research direction of talking face generation in the future.

Keywords　Face generation,Video generation,Image generation,Deep learning,Multimodal learning,Face reconstruction,Deep fake,Computer vision

1　引言

随着数字人、元宇宙等概念的兴起,人体行为生成逐渐成为计算机视觉领域的研究热点。作为其重要分支之一,说话人生成旨在根据输入的多模态数据生成指定对象的说话视频。该任务主要关注于人体的面部区域,目标是生成逼真的人物外观,同时保证说话口型与语音内容的一致性。在实际应用场景中,一个合成的高质量说话人视频需要满足以下

到稿日期:2022-10-07　返修日期:2023-02-20
基金项目:中国国家铁路集团有限公司科技研究开发计划课题(N2021X026)
This work was supported by the China National Railway Group Co., Ltd. Science and Technology Research Project(N2021X026).
通信作者:戴琳琳(daizi2407@163.com)

几个方面的要求：1）清晰程度高且无像素抖动；2）人物口型与语音内容保持一致且面部动作连贯、自然；3）生成视频应当注重头部姿态和相应的面部细节，如张嘴、眨眼等，以增强视频保真度。音视频模态之间的差异以及不同对象在外观上呈现出的多样性，导致说话人生成的主要难点在于保证口型和语音内容的一致性以及相邻视频帧之间面部活动的连贯性。说话人生成在影视传媒、游戏动漫等诸多领域中有着广泛的应用，如短视频中的“换脸”道具、影视中的配音视频生成、动漫游戏中卡通形象的合成等。同时，在近几年热门的虚拟现实（Virtual Reality，VR）、人机交互、元宇宙等新兴应用场景中，说话人生成也逐渐成为不可或缺的一部分。

得益于深度学习的飞速发展，生成式对抗网络[1]（Generative Adversarial Networks，GAN）等深度生成模型的涌现，使得计算机能够在不借助照相机、三维扫描仪等外部设备的情况下，利用深度神经网络模型从输入数据中挖掘人物特征、音频特征等信息，合成逼真的视觉场景。这为说话人生成研究提供了先决条件。在此基础上，目前该领域已经提出一系列说话人生成方法，并且研究热度有逐年增长的趋势。

本文针说话人生成领域的经典工作进行概述和总结，介绍了说话人生成的关键技术，包括三维人脸重建、人脸关键点检测和生成式对抗网络。同时，本文综述了说话人生成的主流方法，并依据其模型分类分别进行介绍和总结。此外，还总结了说话人生成任务中常用的视听数据集和评价指标，分析了现有方法的局限性和提升空间，并对其未来可能的研究方向进行了展望。

2 说话人生成相关技术概述

本节介绍了说话人生成任务中的几种常用技术。1）三维人脸重建的 3DMM 方法能够构建说话人的三维人脸模型。利用音频特征驱动三维模型进行形变，再通过渲染就可以得到二维的生成结果。2）人脸关键点检测技术能够定位包括眉毛、眼睛、鼻子、嘴巴、面部轮廓等面部关键区域的位置。得到关键点后，可以通过进一步操控关键点的位移来模拟说话人的面部动作和口型变化。3）GAN 凭借其生成结果的多样化、真实性，被广泛应用于视觉生成任务。与其他的生成网络相比，GAN 的优势在于能够合成语义丰富的高保真说话人脸图像。

2.1 三维人脸重建

三维人脸重建技术是计算机图形学领域的经典研究方向，旨在建立人脸的参数化模型，并在其基础上改变面部属性，继而渲染出新的人脸图像。实现三维人脸重建的第一步就是要获取人脸的三维模型，其主要方法包括软件建模、仪器采集以及基于图像的建模。其中，软件建模对人力和精力的需求较大，而仪器采集的成本太高，因此基于图像的“低消耗”重建方法得到了广泛研究，其中最典型的方法为三维人脸形变模型（3D Morphable Model，3DMM）。3DMM 的核心思想是将目标的三维人脸模型用数据库中众多人脸的基向量进行线性加权表示。人脸包含形状和纹理两种属性，因此每幅人脸图像可以表示为形状向量和纹理向量的线性叠加。求解一个 3DMM 人脸表示就是估计相应的加权系数。Blanz 等[2]于 1999 年提出了基于主成分分析的经典方法，采用“Analysis-by-Synthesis”的求解思路计算 3DMM 人脸的权重。Paysan 等[3]在 2009 年发布了 BFM 数据集（Basel Face Model），利用激光扫描仪采集了 200 个三维人脸数据，有效地扩大了 3DMM 的使用场景。此外，与 2009 年的版本相比，2017 年的版本 BFM 2017[4]中还增加了人脸的表情参数。

随着深度学习的发展，利用深度学习实现三维人脸重建的方法不断涌现。与传统方法通过优化求解相关系数不同，基于深度学习的方法可以利用模型直接学习相关系数。Tran 等[5]提出了 3DMM CNN，利用 ResNet 网络回归人脸模型的形状和纹理系数。MoFa[6]不依赖于真实数据集，而是将二维图像重建为三维模型后再重新投影回二维图像。这种自监督的方法解决了三维人脸模型数据集获取成本较高的问题。3DDFA[7]和 PRNET[8]等方法基于 3DMM 模型学习人脸的三维特征编码，充分发挥了 CNN 对于像素特征的回归能力。尽管 3DMM 模型的变体层出不穷，但其仍面临诸多挑战：重建的模型受限于人脸区域，缺少眼睛、牙齿及头发等信息；模型的参数空间维度较低，难以重现面部细节。

2.2 人脸关键点检测

人脸二维关键点检测被广泛应用于人脸对齐、表情识别、人脸重建等任务。基于二维面部坐标表示的说话人生成方法通常需要利用人脸关键点检测来得到面部参考坐标，然后采用深度学习模型学习音频特征和参考坐标信息，继而预测说话人面部坐标变化，最后经过生成模型得到视频帧序列。常用的人脸关键点检测方法主要包含 3 类：基于模型的方法、基于级联姿势回归的方法和基于深度学习的方法。

传统的基于模型的代表性方法有 ASM（Active Shape Model）[9]和 AAM（Active Appearance Model）[10]。ASM 利用人工标定预处理的训练集进行训练得到形状模型，再利用搜索策略进行关键点匹配。AAM 在此基础上增加了额外的纹理模型。这两种方法的模型逻辑简单且易于理解，但运算效率较低。级联姿势回归方法 CPR（Cascaded Pose Regression）[11]利用一系列回归器预测形状增量，逐步优化初始形状向量，经过迭代得到最终的关键点结果。DCNN[12]提出了级联的 CNN，采用级联回归思想，首次将 CNN 应用到人脸关键点检测领域。近几年，许多研究团队提出了端到端的人脸关键点检测应用框架，如 OpenCV 的 Dlib 库、OpenFace、PFLD 等，其为使用者提供了便捷、多样化的人脸关键点检测功能。

2.3 生成对抗网络

Ian Goodfellow 等[1]在 2014 年提出的生成对抗网络是一种无监督生成模型，其主要思想是模拟博弈论中的二人零和博弈过程。一个原始的 GAN 网络包含一个生成器 Generator（G）和一个判别器 Discriminator（D）。生成器 G 以随机噪声为输入，在判别器的引导下，不断将噪声拟合成接近真实图像的数据来欺骗判别器 D；而判别器 D 则负责区分生成器 G 的生成结果与真实情况的差异，判断生成结果的“真假”，并将判别结果反馈给生成器 G。因此，GAN 的目标函数如下：

$$\min_G \max_D V(D,G)=\mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)]+\mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{1}$$

其中，$\max_D V(D,G)$ 表示在生成器 G 固定的情况下，判别器 D

通过最大化交叉熵损失 $V(D,G)$ 来更新参数。而 $\min_G \max_D V(D,G)$ 则表示生成器 G 要在判别器判别真、假图片能力最大化的情况下，最小化这个交叉熵损失。在这样的交替训练中，G 和 D 构成了一个动态的博弈过程，并最终达到纳什均衡状态。在实际训练中，生成器和判别器采取交替训练策略，即先训练 D，然后训练 G，不断往复。最终生成器的生成结果能够以假乱真，使判别器的准确率达到 50%。

与其他的生成网络如自回归模型、变分编码器(Variational AutoEncoder，VAE)等模型相比，GAN 能够生成逼真的高质量图像，且应用场景广泛。绝大多数的说话人生成方法都采用 GAN 作为模型框架中的图像生成部分。利用 GAN 模型结构多样化的特点，研究人员可以选择合适的网络构建生成模型，同时根据需求灵活地改变网络输入。例如，DAVS[13]的生成网络包含 10 个卷积层和 6 个双线性上采样层，网络的输入为解耦的音频特征、视觉特征和身份特征。而作为基于二维面部表示的方法，Zakharov 等[14]提出的方法将二维面部坐标作为生成器和鉴别器的输入，并基于 Isola 等[15]的图像-图像转换模型构建生成网络。此外，部分方法基于 GAN 的变体实现说话人生成。Yin 等[16]提出了 StyleHEAT，利用预训练的 StyleGAN 合成高清说话人脸。得益于 StyleGAN 对人物风格可编辑的特点，该方法还支持对人脸属性和视频视角的自由编辑。

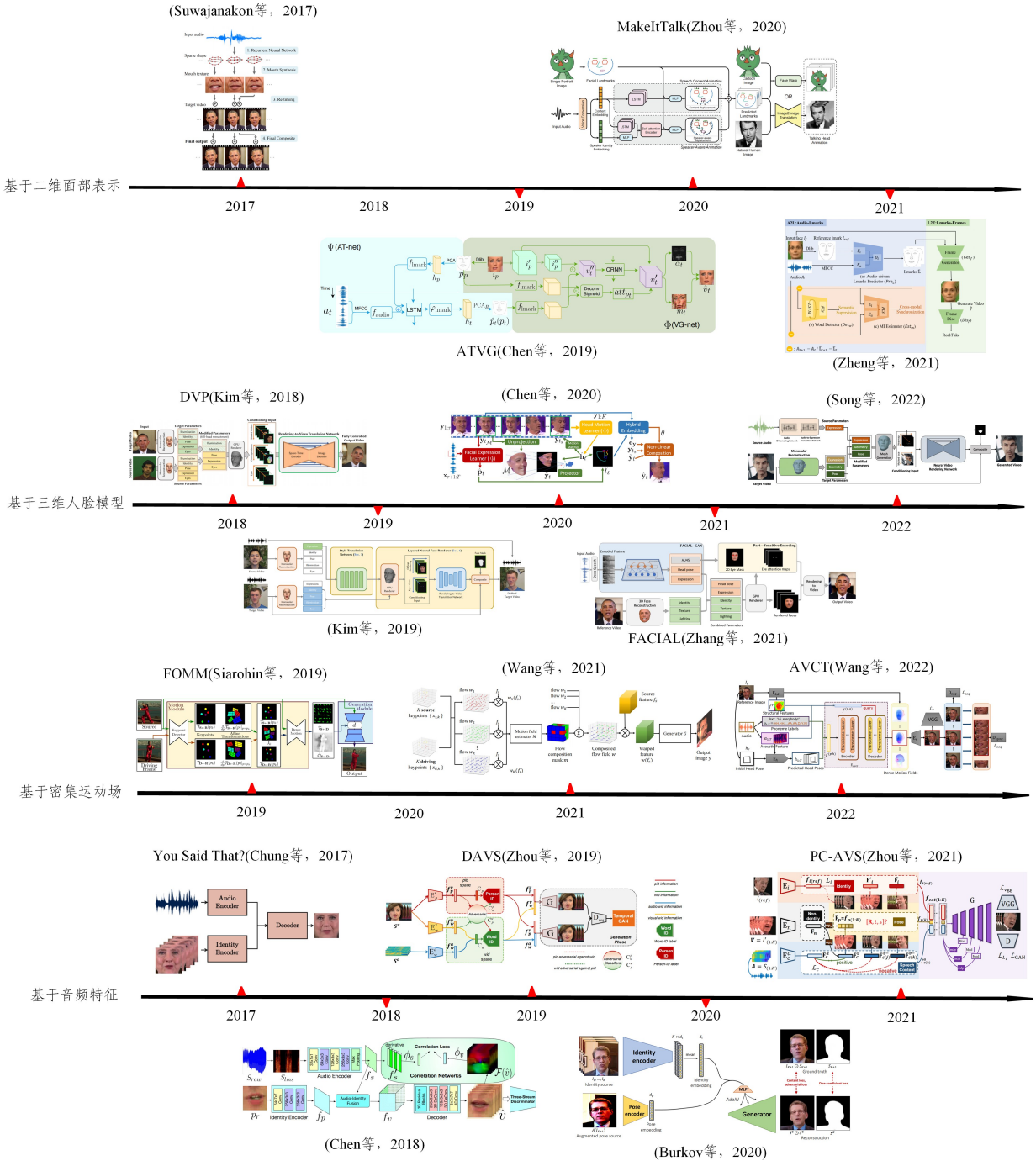

图 1 说话人生成方法的发展历程

Fig. 1 Development of talking face generation methods

3 说话人生成方法及进展

本文根据不同的特征表示类型对说话人生成方法进行分类，几类主流的说话人生成方法的发展历程如图 1 所示。此外现有方法还包括基于视觉特征和基于神经渲染的方法。图 2 给出了说话人生成方法的主要流程。模型基于输入音频提取音频特征，同时从人物图像或视频中提取身份信息，利用不同的特征表示方法，将联合特征作为生成网络的输入，生成对应的说话人视频。本节将结合具体方法分类介绍近年来流行的说话人生成方法和模型结构。

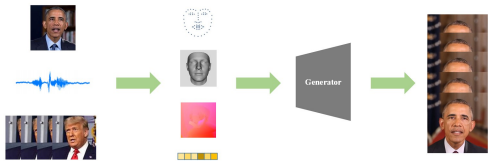

图 2 说话人生成的主要流程

Fig. 2 Basic process of talking face generation

3.1 基于二维面部表示的方法

基于二维面部表示的方法主要以人脸关键点坐标为中间表示，在保留人物身份信息的同时，根据关键点坐标的位移引导生成图像中的面部动作。二维面部坐标的优势在于能够简洁、直观地表达面部动作，但关键点个数的限制导致其表达能力不足，容易缺失面部细节。同时，生成的结果通常仅限于人脸区域，不包含头部姿态及躯干部分。此类方法首先从图像中预测二维人脸坐标，利用音频特征预测说话过程中人脸各部分的坐标变化。由于相邻帧之间的面部坐标具有相关性，因此通常采用长短期记忆网络（Long Short-Term Memory，LSTM）预测二维人脸关键点，模型结合预测坐标与输入图像生成说话人结果。

Suwajanakorn 等[17]提出了说话人脸生成的经典方法 Synthesizing Obama，该方法针对固定对象学习语音到嘴部区域的映射：从音频中提取梅尔频率倒谱系数（Mel-Frequency Cepstral Coefficients，MFCC）音频特征，经过单向 LSTM 的时间延迟循环神经网络（Time-delayed RNN）映射到每一帧对应的稀疏口型轮廓。基于口型关键点的 PCA 特征合成人脸下半区域的加权中值纹理，并与高频的牙齿部分细节相融合，再一并渲染到原始图像中。生成结果保留了原始视频中的背景、躯干以及人脸上半部分图像，仅包含人物的口型变化。

在此基础上，更多的方法关注于完整人脸的二维坐标表示。Chen 等[18]提出了包含两个网络 AT-net 和 VG-net 的 ATVG 框架。该方法首先检测真实图像的面部坐标，并将其作为两个网络的输入。AT-net 根据音频信息与初始的面部坐标，合成说话人的动态面部关键点；VG-net 则根据预测的坐标结果生成视频帧。这两部分网络分别进行独立训练以避免误差的累积，有效提高了说话人口型与语音的一致性。Das 等[19]、Zheng 等[20]和 Zhou 等[21]的方法均采用了类似的级联结构，将模型划分为坐标生成网络与图像生成网络两部分。Das 等[19]在面部坐标生成网络中引入眨眼生成模块，能够得到包含眨眼动作的说话人脸。Zhou 等[21]提出的 MakeItTalk 则从音频输入中同时提取语音特征和身份特征，共同约束面部坐标的位置变化。其中身份特征提供了人物的个性化信息，因此该方法实现了包含头部活动的说话人生成。

Zakharov 等[22]的方法不以音频数据为输入，而是利用图像特征与面部坐标表示直接生成对应姿态下的人物说话图像。此外，该方法采用了大型数据集上的元学习训练方式，能够较好地应用于未知身份的说话人生成。在此基础上，其 2020 年的工作[23]将人物图像分解为低频的姿态粗略图以及包含高频细节的纹理图像，并采用两个生成网络分别生成。其中，面部坐标用于生成低频的人物姿态图像，再与纹理图进行叠加，继而得到具有真实细节的说话人图像。这种组合方式在保证图像视觉质量的条件下，明显加快了生成速度。

Gu 等[24]、Zeng 等[25]和 Zhang 等[26]的方法均采用了包含多个子网络的结构框架，各子网络分别关注说话人视频中的不同信息。例如，Zeng 等[25]的模型中包含坐标编码器、身份编码器、情感编码器、噪声编码器和音频编码器共 5 个编码网络，其分别提取对应的特征向量。解码网络将联合特征转换为不同情绪下的说话人图像，其中说话人的表情与情感特征相对应。与上述直接生成说话人视频的方法不同，Zhang 等[26]的方法利用 4 个卷积神经网络（Convolutional Neural Network，CNN）来恢复深度压缩的说话人图像。这 4 个子网络分别处理音频特征、面部坐标、视频特征和编解码信息，同时引入多个跨模态先验信息，以消除视频中的压缩伪影，从而得到包含高清细节的解压缩结果。

3.2 基于三维人脸模型的方法

考虑到二维人脸关键点有限的表达能力，许多方法着眼于将三维人脸模型作为说话人生成方法的中间表示。基于三维人脸模型的方法通常采用三维人脸重建技术，首先从人脸的二维图像中重建三维人脸模型，得到参数化表示，接着在音频特征的引导下获得说话人的参数化人脸模型，然后经过转换网络渲染得到说话人的二维图像。基于三维人脸模型的典型方法如图 3 所示。

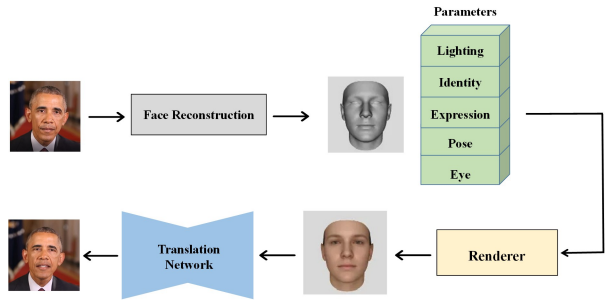

图 3 基于三维人脸重建的说话人生成方法示例

Fig. 3 Example of talking face generation methods based on 3D face reconstruction

利用音频特征精准地操控三维人脸模型，是基于三维人脸模型方法的核心。Richard 等[27]提出了基于语音构建动态人脸三维网格表示的方法，利用自回归采样策略学习音频信号和表情信号在潜在分类空间中的特征表示，并将其统一输入 U-Net 结构的解码网络。该方法采用跨模态损失实现上、下面部区域解耦，通过控制指定的人脸 3D 网格模型来实现逼真的说话动作。Lahiri 等[28]提出了一个从音频中生成个性

化的三维说话人脸的学习框架。该方法首先在一个规范化的空间中解耦人脸的3D几何、头部姿势和纹理，得到音频同步的3D人脸网格。这种规范化将预测问题分解为3D人脸形状和相应的2D纹理图的回归问题。然后，引入自回归纹理预测方法，用于时序稳定的视频合成。此外，作者还提出了一种光照强度规范化算法，以消除视频中的光照变化，使得生成网络能够适应不同环境的光照条件。

Kim等[29]的方法以源视频和目标视频为输入，利用单目人脸重建得到两个视频中人物的三维人脸模型，再分别提取低维参数表示。在参数空间中，用源人物的头部姿势、表情和眼神参数来替换目标人脸参数，以得到低频的人脸渲染图，再经过视频渲染网络得到逼真的目标人物肖像。目标人物的口型和姿态由源视频控制，实现了从源视频到目标视频的人脸重演。在此基础上，作者在后续工作[30]中加入了风格保留网络，以维持目标对象的情感信息，丰富了说话人的面部表情。与Kim等的工作类似，Ren等[31]提出的PIRenderer同样实现了说话人的姿态控制。其不同之处在于，该方法将三维人脸重建提取的面部参数经过映射网络转换为进一步解耦的潜在向量 z，之后再将其输入到变形网络和编辑网络中，生成最终的图像。此外，Ren等还扩展了音频驱动的人脸重演模块，利用LSTM网络，基于已生成的三维人脸模型序列和音频序列来预测说话人对应的三维人脸参数。

与上述人脸重演的方法不同，Chen等[32]的方法利用音频特征驱动三维人脸。该方法引入了面部表情和头部姿态的学习网络，并在生成阶段将嵌入特征进行非线性组合，以减轻大幅度的姿态变化所引起的视觉不连续性，实现了包含自然头部运动和表情的说话人视频生成。Song等[33]、Yi等[34]和Zhang等[35]的方法具有相似的结构，均先从音频特征和三维人脸模型中提取面部几何、表情和头部姿态等属性参数，然后在特征空间中组合得到新的人脸参数化表示，最后经过渲染网络生成具有个性化面部属性的动态说话人结果。

在上述方法的基础上，Ji等[36]的方法同时结合了二维坐标与三维模型这两种中间表示，并引入了情感信息。在第一阶段，该方法首先从音频中利用交叉重建方法解耦情感特征和语音特征，结合人脸关键点来预测说话人的面部坐标。在第二阶段，预测坐标经过一个三维感知的关键点排列模块得到三维人脸属性，之后将重建得到的三维模型映射到人脸区域的轮廓图。渲染网络结合坐标图和输入的参考视频来渲染说话人图像。其生成结果包含个性化的人物属性，同时语音中蕴含的情感信息由说话人的面部表情传达。

3.3 基于密集运动场的方法

基于密集运动场的方法是利用运动流估计人物运动的方法。运动流可以理解为空间中运动物体在观察成像平面上的像素运动的瞬时速度，用于描述物体在视频序列的连续帧之间的运动。此类方法通常需要从输入图像中检测人体关键点，通过预测关键点的移动来计算运动流，从而得到连续帧之间的说话人面部姿态变化和头部活动。基于密集运动场的方法的优势在于通常仅需要目标对象的单张图像作为输入，即可利用音频特征建模密集运动场，其生成结果中普遍包含说话人的头部姿态等属性。

人体密集运动场示意图如图4所示。

图4 人体密集运动场示意图

Fig. 4 Dense motion field map of human body

Siarohin等[37]于2019年提出了Monkey-Net，首次将运动场引入动态人体行为生成任务。该网络由运动关键点检测器、运动估计网络和图像生成网络3部分组成，能够利用参考视频驱动静态图像中的对象完成相应动作。其中，运动估计网络将描述运动变形的稀疏关键点转换为密集运动流，以更好地表达运动信息。作者在后续工作中提出的一阶运动模型(First Order Motion Model，FOMM)[38]利用一阶局部仿射变换扩展了Monkey-Net，能够更好地处理复杂的大幅度姿态变化。上述工作主要关注于人物的整体活动，而非说话人的面部区域。受FOMM的启发，Wang等[39]在其基础上实现了多视角的说话人生成。该方法提取出源图像中人物在标准空间中的三维关键点，利用驱动视频中的关键点运动流操控说话人的姿态和表情变化。同时，该方法采用一种无监督的方法来解耦外观特征与运动特征，从而生成了新视角下的说话人图像。

针对音频驱动的说话人生成任务，Wang等[40-41]的工作引入了音频信息驱动运动场中的关键点移动。其中，Audio2Head方法[40]采用循环神经网络(Recurrent Neural Network，RNN)设计了一个头部姿态预测器，从驱动音频和单张参考图像中估计与语音内容相对应的说话人头部姿态。结合头部姿态和关键点，运动场生成器通过预测关键点的运动来得到密集运动场。图像生成网络以运动场为输入，生成包含自然头部活动的说话人视频。基于关键点的运动场表示集成了面部区域、头部和背景的活动，因此可以更好地约束生成视频的时空一致性。在此基础上，Wang等[41]以音频信号的音素和基于面部关键点的运动场表示为视听关联转换器的输入，将方法扩展到了任意音频输入和任意身份的说话人生成。

部分基于密集运动场的研究结合了三维人脸参数化模型，能够预测更加准确真实的人脸关键点。Zhang等[42]提出的HDTF方法包含两个模块：音频-动画模块和动画-视频模块。其中，音频-动画模块从音频和参考图像中提取人脸参数表示；动画-视频模块借助3DMM估计面部区域的密集运动场，同时用人脸整体的平均移动和面部轮廓的移动分别估计躯干部分与头发区域的运动场，通过组合得到整体的密集流 F^{app}。生成网络在密集流的引导下合成高清的说话人图像，但该方法所展示的生成结果中头部姿态变化不明显。类似的研究还有Doukas等[43]提出的HeadGAN。HeadGAN支持音频和视频两种输入，利用目标三维人脸属性和参考图像计算说话人脸的密集流表示，结合GAN框架的生成网络，能够生成高保真的说话人视频，同时支持对人物表情和姿态的编辑。

3.4 基于音频特征的方法

上述3类方法均利用显式的中间表示作为音频信息与

视觉信息融合的“桥梁”，引导说话人视频生成。相反，基于音频特征的方法不依赖于与人脸相关的显式中间表示，而是直接从输入的音频信号和人物图像中提取语义信息和身份信息作为生成网络的输入。典型的基于音频特征的说话人生成方法的结构如图 5 所示。

图 5 基于音频特征的说话人生成方法示例

Fig. 5 Example of talking face generation methods based on audio feature

基于音频特征的方法大多采用编解码结构的生成网络来预测说话人的口型变化。例如，Chen 等[44]、Chung 等[45]、Prajwal 等[46]和 Vougioukas 等[47]的早期工作利用多个 CNN 组成的编码器提取语音特征和身份特征，再经过解码网络合成口型同步的说话人脸。大部分方法引入多个鉴别器进行训练，以提高生成结果的视觉质量和音画同步性。其中经典的 Speech2Vid 方法[45]的核心思路是学习目标人脸与音频在特征空间上的联合分布。将音频信号和对齐后的人脸图像输入基于 VGG-M 的编码器，编码后的特征经过全连接层得到联合特征，再经过反卷积解码网络重现人脸图像。与上述方法不同，Song 等[48]的方法采用条件循环对抗网络构建生成网络，便于处理视频帧序列之间的时序关系，使说话人的口型变化和面部活动得到平滑过渡。

Zhou 等[13]提出的 DAVS 将提取到的特征经过对抗性训练分类器，解耦视频中的语音信息和身份信息，实现了任意对象的说话人脸生成，但无法生成说话人的头部姿态。作者后续提出的 PC-AVS[49]在其基础上增加了姿态编码器，从数据增强后的输入中提取不包含身份信息的头部姿态特征，同时结合音频特征对姿态特征进行调整，利用隐式模块化视听表示实现了姿态可控的说话人生成。类似地，Burkov 等[50]同样借助姿态编码器引入头部姿态变化。Zhu 等[51]关注音视频信息的跨模态相关性，通过提出的非对称互信息估计器（Asymmetric Mutual Information Estimator，AMIE）估计视听相关性，将音频信息和身份信息在特征空间中融合，生成任意对象的说话人脸视频。此外，作者还设计了一个动态注意力模块，在训练阶段选择性地关注输入图像的唇部区域，进一步提高了口型和语音的同步性。

Zhang 等于 2020 年提出的 APB2FACE[52]和 APB2FACEV2[53]，均以音频特征、头部姿态信息和眨眼信息为联合输入来预测说话人面部几何，再经过生成网络再现人脸图像。两者的区别在于前者通过预测得到二维面部坐标表示；而后者则采用一种更高效的方式，将联合特征经过提出的自适应卷积模块（AdaConv）输入生成网络。这种轻量级的网络结构提高了生成效率，能够实时地在 CPU 或 GPU 上运行。Lu 等[54]的方法同样让实时的说话人生成成为可能。该方法针对任意的音频输入，利用流形投影提取并重构深度语音特征，用于预测目标人物的上半身姿态活动。结合预测姿态的特征图和目标人物参考图像，生成网络能够实时地合成包含头部活动的高保真个性化面部细节。

3.5 基于视觉特征的方法

上述的几类方法大多依赖于音频输入，即从音频信号中提取语音信息，来驱动目标人物的口型变化。与音频驱动的说话人生成方法不同，基于视觉特征的方法通常以视频为输入，通过提取视觉特征，将驱动视频中的说话人口型和姿态转移到目标对象，实现说话人的人脸重演。前文提到的 Burkov 等[50]、Kim 等[29-30]、Ren 等[31]、Zakharov 等[22-23]的方法均为基于视觉特征的方法。

Wang 等[55]于 2018 年提出了经典的 vid2vid 方法，其通过学习从输入语义视频到输出视频的映射函数，使输出视频能够准确描述源视频内容，实现了高分辨率的视频转换。该方法在 GAN 框架中加入光流约束信息，对视频中的前后景分别建模，从多种语义输入中合成逼真、连贯的视频。对于说话人生成任务，vid2vid 能够从提取的人脸轮廓图序列（Edge Map）中合成逼真的说话人视频。鉴于 vid2vid 需要大量场景数据进行训练且泛化能力有限，该团队在其基础上又提出了 Few-shot vid2vid[56]，其引入了注意力机制，仅利用少量示例图像输入即可合成相应的真实场景。

Wang 等[57]、Wiles 等[58]、Wu 等[59]的方法均采用了编-解码器结构，以同时接受源视频和驱动视频输入。X2Face[58]包含嵌入网络和驱动网络两部分，驱动网络以驱动视频帧为输入，将嵌入结果映射到目标图像，以改变目标对象的姿态和表情。ReenactGAN[59]将输入图像编码为人脸轮廓，利用转换网络调整目标对象的面部坐标，而不是在像素空间中直接传递信息，有效避免了生成图像中的伪影现象。LIA[57]是一种自监督的自动编码器，在潜在空间中应用线性运动分解学习一组正交运动方向，利用正交基的线性组合表示潜在空间中的位移向量，预测目标视频中的人物活动。与前几种方法相比，LIA 不依赖于从驱动视频中提取的结构表示，因此在源图像和驱动视频之间存在较大外观变化的情况下能够展现出优良的效果。

Song 等[60]的方法将说话人生成扩展到一种新奇的场景：利用人的说话视频驱动形似人脸的静态物体“说话”，由用户定义的虚拟面孔的“眼睛”和“嘴”跟随驱动视频中的人物同步移动。该方法先分别提取人脸和虚拟面孔的二维关键点，并将其转移到光流图中，然后利用一阶运动近似将关键点的局部运动扩展到虚拟面孔的全局运动场，最后无监督的纹理生成器以原始的静态图像和运动场为参考合成“说话”的虚拟面孔。

3.6 基于神经渲染的方法

与基于三维人脸模型的方法依赖显式表示得到渲染图像的策略不同，基于神经渲染的方法利用神经网络隐式地建模三维人脸，渲染得到二维结果。基于神经辐射场的说话人生成，是目前典型的基于神经渲染的方法。Mildenhall 等提出的神经辐射场（Neural Radiance Field，NeRF）[61]是一种隐式场景表示方法，能够用神经网络隐式地建模一个复杂的三维场景，并通过体渲染（Volume Rendering）从不同视角渲染清晰的场景图片。给定稀疏图像与相机参数，NeRF 学习连续的体辐射场 $F:(\boldsymbol{x},\boldsymbol{d})\rightarrow(c,\sigma)$ 的映射实现场景表示。具体来说，F 是由多层感知器（Multilayer Perceptron，MLP）建模的隐式函数，其输入为三维空间坐标 $\boldsymbol{x}=(x,y,z)$ 与视角方向

$\boldsymbol{d}=(\theta,\varphi)$组成的五维向量，输出为隐式场的颜色 $\boldsymbol{c}=(r,g,b)$ 和体密度 σ。在体渲染阶段，为了计算图像中对应像素的颜色，NeRF 沿射线路径进行数值积分来近似颜色值。考虑从相机中心 $\boldsymbol{o}$ 发出的射线 $\boldsymbol{r}(t)=\boldsymbol{o}+t\boldsymbol{d}$，其预期颜色采用如下方式进行计算：

$$C(\boldsymbol{r})=\int_{t_n}^{t_f} T(t)\sigma(\boldsymbol{r}(t))\boldsymbol{c}(\boldsymbol{r}(t),\boldsymbol{d})\mathrm{d}t \tag{2}$$

其中，$T(t)=\exp\left(-\int_{t_n}^{t}\sigma(\boldsymbol{r}(s))\mathrm{d}s\right)$是射线从 t_n 到 t 这一段路径上的累计透明度。

在此基础上，Gafni 等[62]提出动态神经辐射场，用于人脸外观的动态建模。其引入 3DMM 来估计人物姿态和表情，再经过体渲染合成人脸图像。Guo 等[63]的 AD-NeRF 首次将神经辐射场应用到说话人生成任务。AD-NeRF 将图像分割为背景、人脸和躯干区域，采用两组独立的神经辐射场，分别渲染说话者面部区域与上身躯干，并将其合成最终的结果。该方法从输入图像与音频信号中分别提取姿态特征和音频特征，并将其直接映射到神经辐射场，隐式地建模三维人脸场景表示。隐式渲染技术从参考点向目标场景的各个方向发出射线，采样射线上各点处的颜色与体密度，再渲染得到不同视角下的图像。此外，生成结果还支持对背景和人物姿态的编辑。

与基于 GAN 的方法相比，基于 NeRF 的方法对图像分辨率的限制较小，但体渲染的采样过程需要消耗大量的计算资源，渲染速度较慢，因此此类方法仍然有较大的提升空间。

3.7 小结

说话人生成技术发展至今，相关研究者已提出大量方法，各类方法的特点与优缺点如表 1 所列。可以看出，不同的方法在不同应用场景下展现出了各自的优势，但也同样有很大的提升空间。在未来的说话人生成研究中，研究者可以针对不同方法的优势和局限性开展进一步的研究。

表 1 说话人生成方法的优缺点

Table 1 Advantages and disadvantages of talking face generation methods

说话人生成方法	特点	优点	缺点
基于二维面部表示的方法	利用人脸关键点引导图像生成	简洁、直观地表达面部活动；计算量较小	表达能力不足，易缺失面部细节；仅能合成面部区域
基于三维人脸模型的方法	利用三维人脸重建技术得到参数化人脸模型	表达能力和可操作性较强	人物头部姿态固定，无法处理牙齿、头发等细节
基于密集运动场的方法	利用关键点运动流预测人体活动	建模说话人的整体运动，包含头部姿态等属性	对计算资源的需求较大
基于音频特征的方法	在特征空间中对多模态特征进行融合	不依赖显式的中间表示，生成速度较快	特征的不完全解耦会影响生成效果
基于视觉特征的方法	利用视频驱动，通过对视觉特征的操作来实现人脸重演	不需要提取音频特征	生成结果依赖于对应的驱动视频，应用场景受限
基于神经渲染的方法	利用神经网络建模场景表示	生成结果分辨率高，支持对背景和视角的编辑	渲染速度慢，计算资源消耗大

在前文提到的方法中，基于二维面部表示的方法受到了人脸关键点个数的限制，表达能力不足，容易损失面部细节；同时，此类方法仅能生成说话人的面部区域图像，生成结果缺乏真实性。与上述方法相比，基于三维人脸模型的方法表达能力有所提高，但由于三维人脸重建技术的局限性，此类方法难以处理说话人的头发、牙齿等细节。相比之下，基于密集运动场的方法能够以目标人物的单张图像为输入，实现任意身份的说话人生成；运动场保留了人物的姿态动作，能够生成高保真的说话人结果。但密集运动场需要记录随时间动态变化的关键点运动流，因此对内存占用和显卡计算能力的要求较高。基于语音特征的方法要求对目标特征完全解耦，目前同样可以实现姿态可控的任意对象说话视频的生成。近期的研究方向主要集中于提高模型的生成速度，从而实现实时的说话人生成。除此之外，基于神经渲染技术（如 NeRF）的方法，可以直接利用神经网络建模视觉场景，在说话人生成任务中仍有较大的发展潜力。未来可以尝试利用神经网络建模动态的说话人场景，再通过渲染得到逼真的结果，充分发挥隐式渲染技术强大的场景表达能力。对于此类方法，目前亟待解决的问题是提高模型的渲染速度，以满足人体动态视频生成的需求。

4 常用数据集与评价指标

为了评估说话人生成方法的性能，研究者采集了若干人物视听数据集用于模型的训练和测试。这些数据集包含了不同场景下的人物说话视频片段，如表 2 所列。可以看出，随着模型生成能力的不断提高，数据集由单词、短语的音视频片段逐渐扩展到完整的语句和时间较长的说话视频，数据集的规模也在不断增加。此外，为了满足说话人生成方法对不同应用场景的要求，近期提出的数据集更加关注数据的多样性，视频包含了分辨率、拍摄视角以及视频中人物表情等多种属性的变化。

表 2 人物视听数据集

Table 2 Character audio-visual datasets

数据集	年份	发布机构	来源	数据集特点
GRID	2006	美国谢菲尔德大学	实验室采集	包含符合一定规律的短语，用于早期语音识别任务
LRW	2016	牛津大学视觉几何团队	BBC 广播电视节目	阅读英文单词的视频片段，常用于唇读识别任务
LRS2	2017	牛津大学视觉几何团队	BBC 广播电视节目	自然情景下的语句音视频数据
VoxCeleb	2017，2018	牛津大学视觉几何团队	Youtube	大规模的语句数据集，常作为说话人生成模型的训练数据
MEAD	2020	商汤科技	实验室采集	不同面部表情的多视角人物说话视频，引入情感信息
HDTF	2021	网易伏羲	Youtube	高分辨率的说话人视频，视频的时长和分辨率可以调整

除此之外，在模型评价方面，尽管生成视频的效果更加依赖于视觉上的直观感受，但是为了客观地评估生成结果的质量，研究者也提出了相应的评价指标来衡量图像的视觉质量以及音视频之间的同步性。本节将重点介绍说话人生成研究中适用范围广泛的人物视听数据集和常用的评价指标。

4.1 说话人生成常用数据集

(1)GRID[64](The GRID audiovisual sentence corpus)数据集由美国谢菲尔德大学团队于2006年提出，旨在为语音感知和语音识别等研究提供实验数据。该数据集在实验室环境下录制，包含34名志愿者的34000条语句。每条语句的构成符合以下规律：在“动词”“颜色”“介词”“字母”“数字”和“副词”这6类单词中分别随机挑选一个单词，按照上述顺序排列组成随机短语，每一类单词共51个。

(2)LRW[65](The Oxford-BBC Lip Reading in the Wild Dataset)数据集是由牛津大学视觉几何团队于2016年提出的大规模视听数据集，由1000条包含500个不同单词的语句组成，分别由数百名说话者讲述。该数据集截取自BBC广播电视节目，所有视频长度均为29帧(即1.16 s)。

(3)LRS2[66](The Oxford-BBC Lip Reading Sentences 2 Dataset)数据集是继LRW数据集发布之后，同样由牛津大学视觉几何团队于2017年提出的另一大规模唇读数据集。与LRW类似，该数据集也来源于BBC广播电视节目，收集了开放世界中的无限制数据，共截取超过1000个说话人的将近150000条语句。与LRW相比，LRS2关注于语句任务，不同的单词数多达60000，数据丰富性更高，更加适用于基于深度学习的唇读模型的研究。

(4)VoxCeleb1[67]和VoxCeleb2[68]是牛津大学分别在2017年和2018年发布的两个不重复的大规模开源音视频数据集。该数据集全部采集自YouTube，包含了演讲、访谈、解说等多个视频场景。其中，VoxCeleb1由1251名说话者的超过10万个说话片段组成，VoxCeleb2则扩展到6112个人的100万多条实例。数据集中每个说话视频的平均时长为8.2 s，不包含静音片段，且音频中带有随机的真实噪音。说话人涉及多个种族、职业、年龄，且男女性别比例均衡。数据集自身划分了开发集和测试集。由于视频的真实性和多样性，VoxCeleb被广泛作为说话人识别与说话人生成模型的训练数据。

(5)MEAD[69]数据集(多视图情感视听数据集)由商汤科技于2020年提出，主要针对包含情感表达的说话人生成任务而构建。该数据集包含60名演员在8种不同情绪下的说话视频，每种情绪共有3个不同级别(中性除外)。这些视频在严格控制的环境中从7个不同角度同时录制，总时长为40h，以提供高质量的面部表情细节。此外，多视角的数据也为针对自由视角下的说话人生成研究提供了可靠的数据支持。

(6)HDTF[42](High-definition Talking Face Dataset)数据集是由网易伏羲在2021年提出的高清音视频数据集，包含了分辨率为720P或1080P的非实验室采集视频。该数据集的收集范围是YouTube上近两年的说话人视频，总时长达到15.8h。视频包含了362名不同的人物，并被裁剪为512×512的人脸视频。作者开源了视频处理部分的源码，使用者可以根据需求调整视频的分辨率、时长以及视频帧的宽度和高度。

4.2 常用评价指标

说话人生成的常用评价指标如表3所列。

表3 说话人生成的常用评价指标

Table 3 Evaluation indicators in talking face generation

评价指标	提出时间	介绍
PSNR,SSIM	2004	从像素层面计算生成结果与真实图像相似度，是最常用的图像质量评价方法
FID	2018	计算特征空间中的分布距离，常用于评价GAN的生成结果质量
LPIPS	2018	更接近于观察者直观的视觉感受
LMD	2018	借助唇部关键点计算唇形相似度
SyncNet Confidence	2017	计算特征空间中的对比损失，衡量音视频同步性

(1)PSNR[70](Peak Signal-to-Noise Ratio)和SSIM[70](Structural Similarity Index Measure)分别为峰值信噪比和结构相似性，是评价图像质量最常用的两种指标。在说话人生成研究中，两者常被用来衡量生成图像的视觉质量。通常来说，更高的指标意味着生成图像与真实图像之间的相似程度更高，模糊、伪影现象更少。

峰值信噪比为峰值信号的能量与噪声的平均能量之比。计算时需要首先计算真实图像与生成图像之间的均方误差MSE，PSNR表示为峰值信号能量与均方误差的比值。

$$MSE=\frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j)-K(i,j)]^2 \tag{3}$$

$$PSNR=10\cdot\log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{4}$$

其中，MAX_I为图片可能的最大像素值。

结构相似性是基于生成图片与真实图片之间的亮度、对比度和结构来衡量两者的相似性。整体的表达式可以写为如下形式：

$$SSIM(x,y)=\frac{(2\boldsymbol{\mu}_x\boldsymbol{\mu}_y+c_1)(2\boldsymbol{\sigma}_{xy}+c_2)}{(\boldsymbol{\mu}_x^2+\boldsymbol{\mu}_y^2+c_1)(\boldsymbol{\sigma}_x^2+\boldsymbol{\sigma}_y^2+c_2)} \tag{5}$$

其中，$\boldsymbol{\mu}_x$，$\boldsymbol{\mu}_y$，$\boldsymbol{\sigma}_x$和$\boldsymbol{\sigma}_y$分别为两张图像的均值和方差，$\boldsymbol{\sigma}_{xy}$是协方差，c_1和c_2是用于保持计算稳定的常数。实际计算时，通常在图片上取一个$N\times N$的窗口，不断滑动窗口计算局部SSIM值，最后取平均值作为全局的SSIM。

(2)FID[71](Fréchet Inception Distance)通过计算生成图像与真实图像在特征空间中的距离，来从特征层面评估图像质量，是生成对抗网络最常用的评价指标，因此通常用于评价基于GAN的说话人生成方法的生成结果。FID利用Google的预训练网络Inception Net-V3将输入图像转移到特征空间中，得到对应的2048维向量。

$$FID(x,g)=\boldsymbol{\mu}_x-\boldsymbol{\mu}_g+Tr(\boldsymbol{\Sigma}_x+\boldsymbol{\Sigma}_g-2\sqrt{\boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_g}) \tag{6}$$

其中，$\boldsymbol{\mu}_x$，$\boldsymbol{\Sigma}_x$，$\boldsymbol{\mu}_g$，$\boldsymbol{\Sigma}_g$分别是真实图像和生成图像对应的特征向量集合的均值和协方差矩阵，Tr表示矩阵的迹，根号表示矩阵的平方根运算。越低的FID意味着生成分布与真实图片的分布之间越接近，图像质量越高。

(3)LPIPS[72](Learned Perceptual Image Patch Similarity)是Zhang等提出的“感知相似度”。PSNR和SSIM等传统指标仅从像素层面计算生成图像与真实图像的相似性，与其相比，LPIPS模仿人的视觉感知，更符合视觉上的直观感受。感知相似度的整体计算过程如图6所示。网络F逐层提取输入图像特征，在通道维度中进行单元归一化，得到第l层的结果$\hat{\boldsymbol{y}}^l$，$\hat{\boldsymbol{y}}_0^l\in\mathcal{R}^{H_l\times W_l\times C_l}$。向量$\boldsymbol{w}^l\in\mathcal{R}^{C_l}$用于缩放激活通道，

计算 l_2 距离并在空间维度上取平均值，得到 x 与 x_0 之间的距离 d_0：$d(x,x_0)=\sum_l \frac{1}{H_l W_l}\sum_{h,w} \| w_l \odot (\hat{y}^l_{hw}-\hat{y}^l_{0hw}) \|_2^2$，最后经过映射网络 G 得到感知相似度 h。

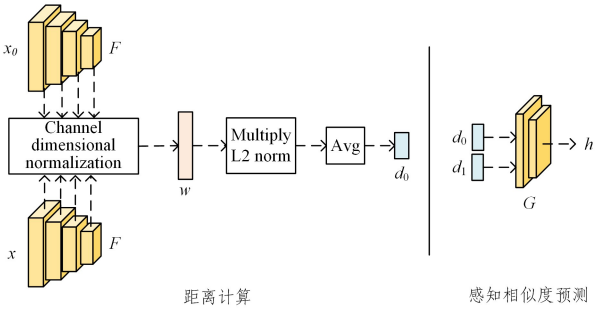

图 6 LPIPS 的计算流程[72]

Fig. 6 Calculation process of LPIPS[72]

(4) LMD[44] (Landmark Distance metric) 由 Chen 等提出。该指标通过计算生成视频与真实视频中的唇部坐标之间的欧氏距离来估计生成口型的准确率，用于衡量生成人物口型的真实程度。计算 LMD 需要标注唇部区域的 20 个稀疏坐标点，并计算对应坐标之间的欧氏距离。但由于坐标点个数的限制，LMD 无法充分检测口型微小的细节变化，其评价的可信程度有待提高。

(5) SyncNet Confidence[65] 是 Chung 等提出的 SyncNet 中的置信度指标。SyncNet 采用双通道网络将音频特征和视觉特征转移到同一参数空间，然后使用对比损失估计特征相似性。

$$E=\frac{1}{2N}\sum_{n=1}^{N}(y_n)d_n^2+(1-y_n)\max(\text{margin}-d_n,0)^2 \quad (7)$$

其中，d_n 为音视频特征向量间的欧氏距离，$d_n=\| v_n-a_n \|_2$。SyncNet 网络使用约 600h 的 BBC 新闻视频进行训练，可以用于计算音视频间的错位时间以及视频中的说话人检测。在说话人生成研究中，研究者常将 SyncNet Confidence 作为衡量音频与视频的同步性的指标，用于描述生成的说话人口型与音频内容的一致性。

5 现有方法的局限性与未来发展趋势

尽管当前的主流方法已经可以生成逼真的音画同步的说话人视频，但是现阶段的说话人生成方法仍存在不少问题。1)生成结果的分辨率与计算资源之间的矛盾。大多数现有方法的生成结果分辨率不超过 512×512，更高的分辨率必然要求更多的计算资源、更高质量的训练数据以及更长的训练时间。2)生成速度较慢。合成一个一分钟的视频片段就需要合成连续的上千张图像，因此说话人生成所消耗的时间通常是普通图像生成任务的几十倍甚至几百倍。3)生成结果视角固定。绝大多数方法只能生成单一视角下的说话人视频，部分支持视角编辑的方法也仅能对观察角度进行微调，大幅度的视角变换常常导致模糊、伪影等现象。4)伪造带来的信息安全问题。随着生成模型性能的不断提升，越来越逼真的生成结果让机器和人眼难辨真假，不法分子可能利用伪造视频传播虚假信息或伪造身份，其可能引发的安全问题不容小觑。

基于对当前说话人生成方法的局限性分析，展望说话人生成研究未来潜在的发展方向：1)在保证生成结果分辨率的前提下优化模型结构，减少计算资源的消耗；2)结合新的高效计算方法，加速图像的渲染和生成过程；3)针对元宇宙、数字人等应用场景需求，未来可以更多地关注三维的说话人场景生成，从任意视角渲染说话人场景；4)合理利用说话人生成算法，协同推进反伪造技术研究的发展。

结束语 说话人生成作为涉及多模态信息的人体行为生成任务，应用场景广泛，具有很高的研究价值。本文结合不同方法分类介绍了当下主流的说话人生成方法的模型结构和方法特点，整理了说话人生成相关的数据集和评价指标。此外，本文还分析了现有方法的局限性，并对说话人生成未来的发展趋势进行了合理展望。

参 考 文 献

[1] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in Neural Information Processing Systems, 2014, 27: 2672-2680.

[2] 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.

[3] PAYSAN P, KNOTHE R, AMBERG B, et al. A 3D face model for pose and illumination invariant face recognition[C]// 2009 sixth IEEE International Conference on Advanced Video and Signal Based Surveillance. IEEE, 2009: 296-301.

[4] GERIG T, MOREL-FORSTER A, BLUMER C, et al. Morphable face models-an open framework[C]// 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition. IEEE, 2018: 75-82.

[5] TRAN A T, HASSNER T, MASI I, et al. Regressing robust and discriminative 3D morphable models with a very deep neural network[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2017: 1493-1502.

[6] TEWARI A, ZOLLHÖFER M, KIM H, et al. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction[C]// The IEEE International Conference on Computer Vision. 2017: 5.

[7] 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.

[8] FENG Y, WU F, SHAO X, et al. Joint 3D face reconstruction and dense alignment with position map regression network[J]. arXiv: 1803.07835, 2018.

[9] COOTES T F, TAYLOR C J, COOPERD H, et al. Active shape models-their training and application[J]. Computer Vision and Image Understanding, 1995, 61(1): 38-59.

[10] COOTES T F, EDWARDS G J, TAYLOR C J. Active appearance models[C]// European Conference on Computer Vision. Berlin: Springer, 1998: 484-498.

[11] DOLLÁR P, WELINDER P, PERONA P. Cascaded pose regression[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2010: 1078-1085.

[12] SUN Y, WANG X, TANG X. Deep convolutional network cascade for facial point detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013: 3476-3483.

[13] ZHOU H, LIU Y, LIU Z, et al. Talking face generation by ad-

versarially disentangled audio-visual representation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019:9299-9306.

[14] ZAKHAROV E,SHYSHEYA A,BURKOVE,et al. Few-shot adversarial learning of realistic neural talking head models[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:9459-9468.

[15] 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.

[16] YIN F,ZHANG Y,CUNX,et al. StyleHEAT:One-shot high-resolution editable talking face generation via pretrained StyleGAN[J]. arXiv:2203.04036,2022.

[17] SUWAJANAKORN S,SEITZ S M,KEMELMACHER-SHLIZERMANI. Synthesizing obama:learning lip sync from audio[J]. ACM Transactions on Graphics,2017,36(4):1-13.

[18] 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.

[19] DAS D,BISWAS S,SINHA S,et al. Speech-driven facial animation using cascaded gans for learning of motion and texture[C]// European Conference on Computer Vision. Cham:Springer, 2020:408-424.

[20] ZHENG A,ZHU F,ZHU H,et al. Talking face generation via learning semantic and temporal synchronous landmarks[C]// 2020 25th International Conference on Pattern Recognition. IEEE,2021:3682-3689.

[21] ZHOU Y,HAN X,SHECHTMANE,et al. Makelttalk:speaker-aware talking-head animation[J]. ACM Transactions on Graphics,2020,39(6):1-15.

[22] ZAKHAROV E,SHYSHEYA A,BURKOVE,et al. Few-shot adversarial learning of realistic neural talking head models[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:9459-9468.

[23] ZAKHAROV E,IVAKHNENKO A,SHYSHEYA A,et al. Fast bi-layer neural synthesis of one-shot realistic head avatars[C]// European Conference on Computer Vision. Cham:Springer, 2020:524-540.

[24] GU K,ZHOU Y,HUANG T. Flnet:Landmark driven fetching and learning network for faithful talking facial animation synthesis[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020:10861-10868.

[25] ZENG D,LIU H,LIN H,et al. Talking face generation with expression-tailored generative adversarial network[C]// Proceedings of the 28th ACM International Conference on Multimedia. 2020:1716-1724.

[26] ZHANG X,WU X. Multi-modality deep restoration of extremely compressed face videos[J]. arXiv:2107.05548,2021.

[27] RICHARD A,ZOLLHÖFER M,WEN Y,et al. Meshtalk:3d face animation from speech using cross-modality disentanglement[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:1173-1182.

[28] 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.

[29] KIM H,GARRIDO P,TEWARIA,et al. Deep video portraits[J]. ACM Transactions on Graphics,2018,37(4):1-14.

[30] KIM H,ELGHARIB M,ZOLLHÖFER M,et al. Neural style-preserving visual dubbing[J]. ACM Transactions on Graphics,2019,38(6):1-13.

[31] REN Y,LI G,CHEN Y,et al. PIRenderer:Controllable portrait image generation via semantic neural rendering[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:13759-13768.

[32] CHEN L,CUI G,LIU C,et al. Talking-head generation with rhythmic head motion[C]// European Conference on Computer Vision. Cham:Springer,2020:35-51.

[33] SONG L,WU W,QIAN C,et al. Everybody's talkin':Let me talk as you want[J]. IEEE Transactions on Information Forensics and Security,2022,17:585-598.

[34] YI R,YE Z,ZHANG J,et al. Audio-driven talking face video generation with learning-based personalized head pose[J]. arXiv:2002.10137,2020.

[35] 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.

[36] 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.

[37] SIAROHIN A,LATHUILIÈRE S,TULYAKOV S,et al. Animating arbitrary objects via deep motion transfer[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:2377-2386.

[38] SIAROHIN A,LATHUILIÈRE S,TULYAKOV S,et al. First order motion model for image animation[J]. Advances in Neural Information Processing Systems,2019,32:7137-7147.

[39] WANG T C,MALLYA A,LIU M Y. One-shot free-view neural talking-head synthesis for video conferencing[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:10039-10049.

[40] WANG S,LI L,DING Y,et al. Audio2Head:Audio-driven one-shot talking-head generation with natural head motion[J]. arXiv:2107.09293,2021.

[41] WANG S,LI L,DING Y,et al. One-shot talking face generation from single-speaker audio-visual correlation learning[J]. arXiv:2112.02749,2021.

[42] ZHANG Z,LI L,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. 2021:3661-3670.

[43] DOUKAS M C,ZAFEIRIOU S,SHARMANSKA V. HeadGAN:One-shot neural Head synthesis and editing[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:14398-14407.

[44] 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.

[45] JAMALUDIN A, CHUNG J S, ZISSERMANA. You said that: Synthesising talking faces from audio[J]. International Journal of Computer Vision, 2019, 127(11): 1767-1779.
[46] PRAJWAL K R, MUKHOPADHYAY R, NAMBOODIRIV 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.
[47] VOUGIOUKAS K, PETRIDIS S, PANTIC M. Realistic speech-driven facial animation with gans[J]. International Journal of Computer Vision, 2020, 128(5): 1398-1413.
[48] SONG Y, ZHU J, LI D, et al. Talking face generation by conditional recurrent adversarial network[J]. arXiv: 1804. 04786, 2018.
[49] 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.
[50] BURKOV E, PASECHNIK I, GRIGOREV A, et al. Neural head reenactment with latent pose descriptors[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 13786-13795.
[51] ZHU H, HUANG H, LI Y, et al. Arbitrary talking face generation via attentional audio-visual coherence learning[C]// Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. 2021: 2362-2368.
[52] ZHANG J, LIU L, XUE Z, et al. Apb2face: audio-guided face reenactment with auxiliary pose and blink signals[C]// IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP 2020). IEEE, 2020: 4402-4406.
[53] ZHANG J, ZENG X, XU C, et al. APB2FaceV2: Real-time audio-guided multi-face reenactment[J]. arXiv: 2010. 13017, 2020.
[54] LU Y, CHAI J, CAO X. Live speech portraits: real-time photo-realistic talking-head animation[J]. ACM Transactions on Graphics, 2021, 40(6): 1-17.
[55] WANG T C, LIU M Y, ZHU J Y, et al. Video-to-video synthesis [J]. Advances in Neural Information Processing Systems, 2018, 31: 1144-1156.
[56] WANG T C, LIU M Y, TAO A, et al. Few-shot video-to-video synthesis[J]. Advances in Neural Information Processing Systems, 2019, 32: 5013-5024.
[57] WANG Y, YANG D, BREMOND F, et al. Latent image animator: learning to animate images via latent space navigation[C]// International Conference on Learning Representations. 2021.
[58] WILES O, KOEPKE A, ZISSERMAN A. X2face: A network for controlling face generation using images, audio, and pose codes [C]// Proceedings of the European Conference on Computer Vision. 2018: 670-686.
[59] WU W, ZHANG Y, LIC, et al. Reenactgan: Learning to reenact faces via boundary transfer[C]// Proceedings of the European Conference on Computer Vision. 2018: 603-619.
[60] SONG L, WU W, FU C, et al. Everything's talkin': Pareidolia face reenactment[J]. arXiv: 2104. 03061, 2021.
[61] MILDENHALL B, SRINIVASAN P P, TANCIKM, et al. Nerf: Representing scenes as neural radiance fields for view synthesis [C]// European Conference on Computer Vision. Cham: Springer, 2020: 405-421.
[62] GAFNI G, THIES J, ZOLLHOFER M, et al. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 8649-8658.
[63] 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.
[64] ALGHAMDI N, MADDOCK S, MARXER R, et al. A corpus of audio-visual Lombard speech with frontal and profile views[J]. The Journal of the Acoustical Society of America, 2018, 143(6): EL523-EL529.
[65] CHUNG J S, ZISSERMAN A. Lip reading in the wild[C]// Asian Conference on Computer Vision. Cham: Springer, 2016: 87-103.
[66] CHUNG J S, SENIOR A, VINYALS O, et al. Lip reading sentences in the wild[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2017: 3444-3453.
[67] NAGRANI A, CHUNG J S, ZISSERMAN A. Voxceleb: a large-scale speaker identification dataset[J]. arXiv: 1706. 08612, 2017.
[68] CHUNG J S, NAGRANI A, ZISSERMAN A. Voxceleb2: Deep speaker recognition[J]. arXiv: 1806. 05622, 2018.
[69] 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.
[70] 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.
[71] 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: 6626-6637.
[72] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 586-595.

SONG Xinyang, born in 2000, Ph.D candidate. His main research interests include human behavior generation and so on.

DAI Linlin, born in 1983, postgraduate, senior engineer. Her main research interests include computer vision and technology and so on.

(责任编辑:柯颖)