数字说话人脸生成技术综述

张冰源[1,2]，张旭龙[1]，王健宗[1]，程宁[1]，肖京[1]
1. 平安科技（深圳）有限公司，广东 深圳 518000；
2. 中国科学技术大学先进技术研究院，安徽 合肥 230000

摘要
在现代计算机视觉和自然语言处理的交叉领域，数字说话人脸生成技术已经成为一个越来越重要的研究主题。数字说话人脸生成技术专注于依据预定的文本或音频序列生成逼真的人脸图像。近年来，深度学习方法，如卷积神经网络、生成对抗性网络以及神经渲染场在此领域已经表现出了显著的应用价值。这些方法不仅引起了学术界的广泛关注，而且在工业界得以实际应用，用于解决图像处理和计算机视觉方面的具体问题。尽管已经取得了一定的进展，实际应用这些方法仍然面临诸多挑战。综合分析和评估深度学习方法在数字说话人脸生成方面的具体实现，以识别现存方法的优缺点，探讨尚待解决的普遍问题，并突出仍需进一步研究的开放性问题。此外，从统计学角度列出了目前可用的数据集，并对其进行评估和比较，以便研究人员能更容易地选择满足他们需求的数据集。

关键词
数字说话人脸生成；虚拟人；语音驱动

中图分类号：TP37 文献标志码：A doi: 10.11959/j.issn.2096-0271.2024059

Survey of audio-driven talking face generation technology

ZHANG Bingyuan[1,2], ZHANG Xulong[1], WANG Jianzong[1], CHENG Ning[1], XIAO Jing[1]
1. Ping An Technology (Shenzhen) Co., Ltd., Shenzhen 518000, China
2. Institute of Advanced Technology, University of Science and Technology of China, Hefei 230000, China

Abstract
In the interdisciplinary field of modern computer vision and natural language processing, digital talking facial generation technology has become an increasingly important research topic. Digital facial generation technology focuses on generating realistic facial images based on predetermined text or audio sequences. In recent years, deep learning methods such as convolutional neural networks, generative adversarial networks, and neural rendering fields have been used for digital talking face generation, which shows significant research and application value. These methods have not only attract widespread attention from the academic community, but also have been applied in industry to solve specific problems in image processing and computer vision. Although some progress has been made, the practical application of these technologies still faces many challenges. Comprehensively review and evaluate the specific implementation of deep learning methods in the generation of digital talking face to identify the advantages and disadvantages of existing methods,

explore common problems that need to be solved, and highlight open issues that still require further research. In addition, currently available datasets from a statistical perspective were listed, evaluated and compared so that researchers can more easily choose datasets that meet their needs.

Key words
digital talking face generation, virtual human, audio-driven

0 引言

在人工智能与多媒体计算的交叉领域，实现虚拟世界与真实世界的自然、无缝交互已经成为一个长期和核心的研究目标。在复杂的计算生态系统中，视觉与听觉信息的高效整合为用户提供了更加逼真和沉浸式的体验，是当前研究关注的热点。因此，数字说话人脸生成（digital talking face generation）技术应运而生，提供了解决这一问题的全新可能。不同于传统的面部生成或面部动画技术，数字说话人脸生成技术不仅要求生成高度逼真的动态人脸图像，还要求能与给定的语音或文本信息进行精准的时序同步。这意味着，在评价该技术的成功性时，除了图像质量，时序一致性和动态表达的能力也是不可忽视的关键因素。

近年来，一系列深度学习方法，如卷积神经网络（convolutional neural network，CNN）、生成对抗网络（generative adversarial network，GAN）以及神经渲染场（neural rendering fields，NeRF）[1-2]等，在数字说话人脸生成方面取得了引人注目的成果。这些方法不仅得到了学术界的高度关注，同时也在工业应用中表现出了巨大的潜力，应用场景包括增强现实（augmented reality，AR）、虚拟现实（virtual reality，VR）、电子游戏、远程会议、虚拟试衣、电影和电视特效、社交媒体上的个性化化身等。然而，如何将这些高度复杂的方法从理论推广到实际应用，特别是在噪声环境和多样化输入条件下保持高性能，依然是待解决的问题。

在此背景下，数字说话人脸生成技术的研究从早期的基础模型探索逐渐转向综合性和多学科交叉的模型研究，涵盖了从基础科学、计算机视觉、自然语言处理到社会伦理和法律问题等多个层面。因模型的多样性、数据集的复杂性、评估准则的不统一性以及应用场景的广泛性，研究者和从业者面临诸多挑战。

1 人脸生成技术

从单个野生图像中生成说话人脸是一个具有挑战性的问题，可以使用未经训练的照片自动实现或使用用户交互界面建模。在介绍较先进的方法之前，先介绍一下人脸生成技术的一些基本的工具：3D可变形模型（3D morphable model，3DMM）、基于CNN和GAN的人脸生成方法。

1.1 3D可变形模型

3DMM[3-5]是实体类别中3D纹理[6-7]和形状[8-9]解释连续参数化的结果。3DMM将低维参数映射到纹理化3DMM模型的高维参数。它以密度函数的形式包含关于实体类的统计数据。3DMM被用于人脸动画和人的音唇同步[10-11]。其他领域也可能会受益于这些形状[12-13]和使用这些方法展示的纹理知识[14-16]。传统的3D形式重建技

术中效果不佳的是3D绘画的重建[14-16]。然而，受益于3DMM技术，这种重建对娱乐目标变得越来越重要。未来的应用程序可能用于检查同一张人脸的不同肖像之间的纹理，模拟空间的形状和外观，并有助于提高3D的人脸重建和地标定位性能。

3DMM探索了一个UV映射[17-18]来展现一个3D的纹理[19]。UV映射将3D纹理分配给2D平面，所有纹理都普遍按像素对齐。它通常是通过圆柱形将平均形状展开成一个2D平面的空间公式，并用于创建一个RGB图像。来自3D区域的每个顶点在存储纹理信息的UV平面中都有一个相关的纹理坐标。PCA（principal components analysis）定义统计纹理表示，所有训练纹理UV映射被向量化。在3DMM研究开始时，纹理的统计模型是由在实验室中捕获的人脸构建的。由于这样的模型不能代表在野外捕获的人脸，研究人员提出了SIFT（scale-invariant feature transform）[20]和HoG（histogram of oriented gradient）[16,21-22]，直接使用野外人脸构建模型。

表面法线的可用性和模拟照明效果是3D面模型的优点。因此，一些作品将3DMM与手机建模的照明模型或球面谐波模型相结合，以生成更真实的人脸图像。3D人脸模型被分为3D形状模型（3D shape model，3DSHM）、3DMM（3D morphable model）和扩展的E3DMM（extended 3D morphable model，E3DMM）。3DSHM模型显式地建模姿态，并可以处理姿态和场景照明文献[13,23]引入深度神经网络重建非线性3D人脸变形模型，以弥补线性模型的缺点，这是一种更灵活的表示。文献[24]利用U3DMM（unified 3D morphable model）来模拟更多的个人内部变异，如咬合。

1.2 使用CNN生成人脸

CNN是一种强有力的计算机视觉的深度学习工具，它在图片处理、目标检测和高质量的合成方面都有广泛的应用。文献[25]提出了一种解决基于特定选择的面部特征合成高分辨率人脸图像问题的方法。一个预先训练的CNN被用于创建人类面部标签的分类系统。此外，一种新的建模技术定制的高斯混合模型（customized Gaussian mixture model，CGMM）被使用在通过CNN分布特征激活的模型中[26-27]。与其余的深度生成方法相比，该方法的主要优点是，除CNN之外，它不需要任何深度神经网络体系结构，根据各种人脸面部特征对人脸图像进行分类，并根据给定的数据合成新的人脸图像。与传统的深层生成技术相比，CGMM使用这种策略有两个主要动机：一是这是一种目前未在CNN图像处理中采用的新技术；二是它不需要复杂的设计来补充当前CNN。因此，所提出的方法比以前的方法更简单，训练的参数更少。实验结果表明，CGMM与人脸特征具有较高的匹配度，并证明了CGMM在生成高质量人脸方面的意义。引入的人脸生成方法有许多可想象的应用，如在执法部门和其他更通用的生产环境中识别嫌疑人。

1.3 使用GAN生成人脸

近年来，将新颖的GAN和对抗训练过程应用到人脸合成过程中取得了优异的成绩。它们主要用于合成中性表情及正面、逼真的人脸照片，人脸识别，以及会说话的人脸合成[28-29]。文献[30-31]对人类面部表情的生成，引入了条件差分对抗自编码器（conditional difference

adversarial auto-encoder, CDAAE)，CDAAE通过合成目标情感图像或人脸动作单元生成人脸图像。为了解决这个问题，他们建议在自动编码器和解码器中添加前馈路径，提出的框架将面部表情与动作进行组合和插值，以创建独特的表情。实验结果表明，所提出的框架在保留人类身份细节的同时，比以前的方法更忠实地合成了未被注意的对象的面部表情。文献[32]介绍了从地标点生成人脸的方法，在感知损失和一种新的性别保留损失的指导下，使用GAN在保留性别信息的情况下合成人脸。研究人员根据文献[33]的转换后的语音创建了会说话的面孔，他们创建了LipGAN，一个新的生成式对抗网络，它使用一个鉴别器测量生成器帧中唇同步的程度。它由两个网络组成：①*G*生成器，它根据输入的音频生成人脸；②*D*鉴别器，它确定生成的人脸是否与输入的音频匹配。*G*生成器在训练过程中准确地生成逼真的人脸，并与音频同步。

RankGAN[34]是一种新的基于边缘的损失函数。如果简单的状态被很好地分类，它允许鉴别器关注有问题的状态，因此分类器变得更加有效。鉴别器*D*通过评估几种生成器类型*G*有效地学习训练数据中的细微差别。框架由一个判别器组成，该判别器从生成器的几个阶段对生成图像的质量进行排序。排名器引导生成器学习训练数据中的细微差别，并在每个阶段逐步改进。

最近的研究使用GAN来丰富输入的3D人脸几何构造。基于纹理映射和粗网格，文献[35]开发了具有精细尺度属性的粗网格。文献[36-37]从输入数据和基本网格中收集了详细的几何形状和高质量的反射率。这两项研究都针对输入数据进行调节，与之前的研究不同的是，它们没有产生生成的3D人脸模型。

一些研究工作已经提出将3DMM与通过对抗性学习获得的印象模型相结合。在文献[38]中，为了合成准确的合成数据，在对齐的面部纹理上训练解耦的GAN，并通过线性3DMM进行连接，训练了一个类似的模型，证明可以用GAN精确拟合2D图像中的纹理信息。作者使用3DMM对图像进行拟合，使用GAN来填补所得UV映射中的空白，其依靠3DMM以有限的表现力来塑造空间。

文献[39]使用了新颖的GAN来学习3D面部形状变化，通过在单个野生图像上训练GAN，演示了学习身份变化。他们主要关注的是面部表情，因此，他们没有模拟由于表情或身份与表情之间的相关性而产生的非线性变化。文献[40]提出了新颖的框架3DFaceGAN，该方法可以处理多标签标注数据。文献[41]通过将相关的形状和纹理转换为矢量，导出了一个3D可变形人脸模型。The Face Ware House[42]系统通过对广泛收集的4D面部扫描数据（这些数据是由RGB-D传感器捕获的）进行多元线性分析，增强了原始的面部扫描数据的中性人脸模型的性能。文献[43]提出了一种从单个野生图像生成3D面部头像的方法，利用深度神经网络直接从给定图像中预测3D人脸模型的顶点坐标。文献[17]提出了一个新颖的UV-GAN（generative adversarial network for UV completion）来实现一个脸部的UV映射并且恢复自闭塞区域。在实验过程中，通过将完成的UV映射附加到拟合的3D面网格上，生成任意位姿下的虚拟实例。特征融合生成对抗网络（feature fusion generative adversarial network，FF-GAN）[44]和双路径生成对抗网络（two-pathway generative adversarial network，TP-GAN）[45]提出了两种基于GAN的人脸正面化方法。FF-GAN将3DMM合并到GAN中，提供形状

和外观先验，在训练数据较少的情况下快速收敛。TP-GAN是通过同时感知全局结构和局部细节的真实感正面视图合成的。文献[46]使用一个实用的计算框架，基于GAN，从语音输入中引入人脸。该架构将生成的人脸身份与训练数据集中说话人的身份相匹配，根据输入的声音生成人脸。文献[47]提出了一种名为AvatarMe的方法。AvatarMe从单个野生图像中重建高分辨率逼真的3D人脸。为了实现人体皮肤的逼真渲染，AvatarMe分别对所需几何形状的漫反射、镜面反照率和法线进行了建模，同时，从输入图像中推断出面部几何形状。

使用GAN可生成更高质量的3D参数来代替3DMM，但统计模型仍然保持维度不变。通常采用带变分自编码器以及GAN的卷积网络来处理UV映射参数。然而，这些方法忽略了形状、法线和纹理之间的相关性，这对于身份空间的真实感至关重要。这种相关性的重要性在不相容的面部属性中表现最明显。文献[48]提出了多路转移的GAN框架保存不同3D模态之间的相关性来解决这个问题，并提出通过将纹理、形状和法线联合，训练一种新的基于树干-分支的GAN（trunk-branch based generative adversarial network，TB-GAN）来建模和合成连贯的3D人脸。该模型旨在保持相关性，同时容忍特定领域在形状、纹理和法线方面的差异，并且可以很容易地扩展到其他领域。Gecer等[48]声称，这是第一个将法线作为额外的信息来源进行通用面部建模的方法，这是对具有身份几何信息的面部生成表达模型、纹理和法线空间表达模型的首次研究。文献[49]提出了一种基于DCGAN（deep convolution generative adversarial network）的方法直接生成3D人脸，被提出的3D人脸模型的几何采样方法为输入的3D人脸输出一个结构化表示。

文献[50]发现，尽管它们具有分层卷积性质，但典型的生成对抗网络的合成过程以一种不安全的方式依赖于绝对像素坐标。例如，表现为与图片坐标绑定的细节查找，而不是表示对象的表面。Karras等[50]将生成器网络中的混叠现象归因于信号处理的疏忽。此外，通过将网络中的所有信号视为连续的，他们得出了广泛适用的、简单的架构更改，以确保不需要的信息不会渗入分层合成过程。生成的网络具有与SOTA（state-of-the-art）相同的FID（fréchet inception distance），但是有非常不同的内部表征，它们与平移和旋转完全相等，即使在亚像素尺度上也是如此。这种方法为图像、动画以及视频生成模型提供了基础。文献[51]在目标模型的基础上建立了他们的模型，并引入了一个学习仿射层，用于输出输入的傅里叶特征的全局迁移和旋转参数。该层被初始化为执行身份转换，但随着时间的推移，在有益的情况下才学习使用该机制。

2 数字说话人脸生成技术

近年来，从语音中自动生成说话人脸的技术引起了研究人员的极大关注。基于人脸图像和驱动源（多个模态的信息）的动态说话人脸生成技术旨在合成与驱动源相对应的逼真的动画说话人脸视频，说话驱动的人脸生成的流程如**图1**所示。这个问题的解决方案对于实现广泛的实际应用至关重要，例如用其他语言重新为视频配音、视频会议或角色扮演视频游戏的远程呈现、带宽有限的视频转换和虚拟锚。其他可能的应用是增强语言理解，同时保护失聪和重听人的隐私，同时也有利于安全领域对抗性攻击的研究，为监督学习方法提供更多的训练样本。然而，由于以下两个原

因，研究动态人脸生成是具有挑战性的。

● 说话人脸的变形是由个体内在的主体特征、外在的镜头位置、头部运动和面部表情组成的，这些特征是错综复杂的。这种复杂性不仅源于面部区域的建模，还源于头部运动和背景的建模。

● 明确地利用参考视频中包含的视觉信息仍是一项挑战。因为在基于学习的方法中很难避免人们在合成视频中出现敏感的微妙伪影和感知身份变化。

数字说话人脸生成技术引起了研究者的广泛关注。该领域的相关研究可分为3类：语音驱动的数字说话人脸生成技术、文本驱动的数字说话人脸生成技术和表现驱动的数字说话人脸生成技术。

(1) 语音驱动的数字说话人脸生成技术

语音驱动的数字说话人脸生成技术旨在通过分析输入的语音信号，自动生成与之同步的逼真的人脸动画，使数字角色能够像真人一样进行自然的交流。首先，对输入的语音信号进行分析，提取其中的关键特征，如音高、音强、音色等。这些特征将作为后续生成人脸动画的重要依据，建立语音特征与面部表情之间的映射关系模型。该模型可以基于大量的人脸表情数据进行训练，学习不同语音特征与相应面部表情之间的关联。根据提取的语音特征和映射模型，生成与语音同步的面部表情序列。这通常涉及对人脸模型进行精细的控制，使其能够准确地表达各种情感和语义。唇形同步是语音驱动人脸生成中的一项关键技术，它需要精确地模拟人在发音时的嘴唇动作，以确保人脸动画的真实性和可信度。

(2) 文本驱动的数字说话人脸生成技术

文本驱动的人脸生成技术已经被开发出来，用于由文本或语音驱动的人脸模型构建。文本驱动的方法通常由文本到语音和文本到人脸形状合成单元组合而成，以

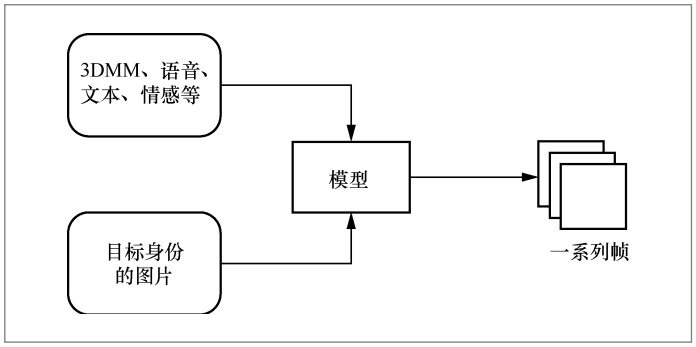

图1 说话驱动的人脸生成的流程

生成动态人脸。

(3) 表现驱动的数字说话人脸生成技术

面部表现捕捉和基于表现的面部生成是较热点的研究领域，许多不同的采集系统和处理管道共享基本原理和具体实施细节。基于表现的人脸生成技术通常包括一个非刚性跟踪阶段和一个表情重定向过程。

2.1 使用CNN的数字说话人脸生成技术

近年来，随着GAN的成功，图像合成的发展取得了长足的进步，这与视觉语音合成的难度密切相关。基于序列像素预测的PixelRNN[52]和PixelCNN[53]架构是有效的视觉语音合成方法，这个架构被条件PixelCNN扩展，可以在任何向量上被条件化。

级联细化网络使用逐像素的语义布局来创建逼真的图像。它采用了先前用于图像样式转移作品的“内容表征”损失机制，由于损失，网络被迫在生成的图像和实际场景之间匹配预训练的CNN的激活，与图像空间网络相比显示出明显的优势。

文献[54]提出了一种基于CNN的说话人脸视频制作方法。该方法将接收到的目标人脸的照片和音频作为输入，之后创建目标人脸与音频对口型的视频。该方法具有实时性，适用于训练期间不可见的音频

和人脸。说话人脸是使用编码器–解码器的聚合体模型生成的，该模型将人脸和声音联合嵌入，使用一种跨模态的自监督方法对未标记的视频数据进行训练。一个多流的CNN模型建议将生成的人脸融合到原始视频的帧中，从而在视觉上重新为视频配音。

文献[55]提出了X2Face网络，可以使用驱动框架中的另一个驱动脸控制源脸，以产生具有源帧身份但具有驱动框架中驱动脸的姿态和表情的生成帧。该网络使用广泛收集来的视频数据进行训练以实现完全自我监督，生成过程可以由其他模态驱动，例如音频或姿势编码，而无须对网络进行任何进一步的训练。Wiles等[55]认为，该网络对输入数据的假设较少，因此它比其他方法更稳健。该网络采用自监督模式进行训练，X2Face的训练过程如**图2**所示。它有3个源和一个驱动架。源的一种形式被指定为源帧，将选择的源帧输入嵌入网络中，采样器将学习像素从源帧映射到嵌入帧中。驱动框架被输入驱动网络中，并学习如何将嵌入人脸的像素映射到生成的人脸。生成的帧应该具有源帧的身份和驱动帧的姿态/表情。由于本训练阶段的帧来自同一视频，因此生成的帧和驱动的帧应该相匹配。在测试时，源的身份和驱动脸可以不同。

2.2 使用GAN的数字说话人脸生成技术

文献[56]介绍了一种基于无声视频的语音重构方法，从人的面部动作中推断语音。该方法遵循编码器–解码器架构，它由两个被称为塔的CNN分支组成。其中，第一个CNN分支输入裁剪后的人脸灰度图像，第二个CNN分支输入给定序列帧的光流。嵌入是通过连接每个塔的输出来创建的。该方法的解码器由输出梅尔尺度频谱图的全连接层和输出线性尺度频谱图的后处理网络设计。该方法的技术是给定图像帧序列的语音推理的一个例子。然而，最新的研究表明，给定一个语音框架，就有可能生成逼真的图像。文献[57]是较早提出解决这一问题的研究之一，其介绍了一种生成人脸对话视频的方法。该方法利用目标语音音频的梅尔频率倒谱系数和一张身份图像输出一个与语音同步的图像。

通过面部表情来传达人脸的语义包括的范围很大，它包括动作、说话、眨眼和情绪状态，比如开心、悲伤或者惊讶。多年来，人们对面部表情编辑进行了大量研

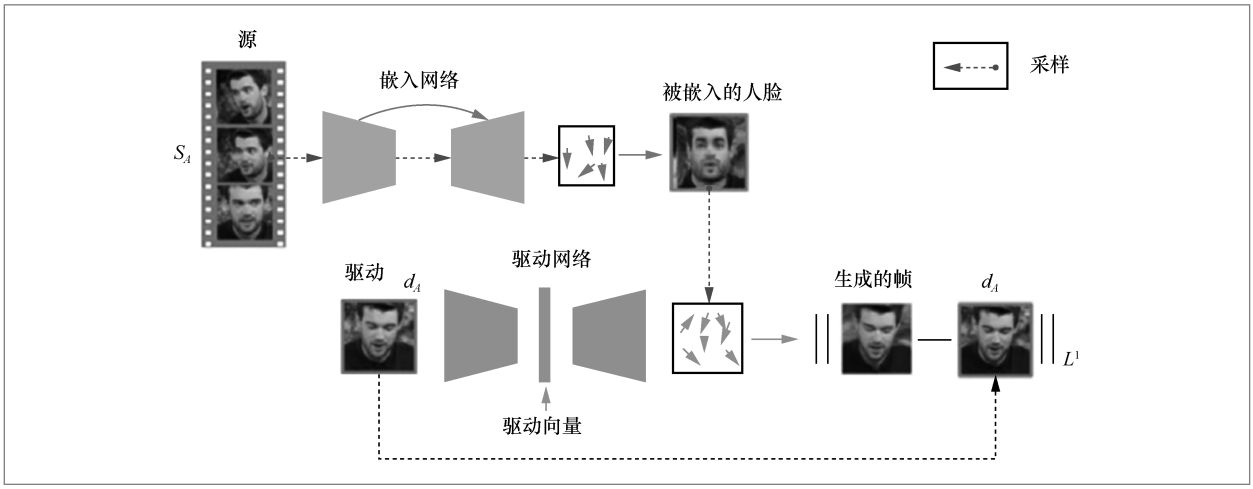

图 2 X2Face 的训练过程

究，将目标面部的语义表达转移到源面部的语义表达，并取得了一定的研究成果。这些先进的方法假设有一对源目标图像可用，并且两个图像中都存在一对相匹配的2D或3D面部网格，用于纹理翘曲和渲染。此外，笔者认为有一组源映像来学习，可以在运行时为源实例创建统计相关特征表示是很重要的。但上述方法将这些技术的应用限制在源数据的某些设置中。为了解决该问题，文献[58]对中级面部表情操作进行了研究，仅通过给定的AU（action unit）系数直接使人的肖像动起来，从而使面部表情编辑任务的灵活性达到一个全新的水平。

在通过语言的方式表达中，不同的情绪可以直接影响沟通中传递的内容信息，改变语言中的情绪甚至可以使其语义发生巨大变化。研究表明，对于未经训练的人来说，纯粹从语音音频中预测情绪是相当困难的。由于人们在情绪解释中严重依赖视觉线索，因此为了使视觉渲染更加逼真，提高语音交流能力，自动说话人脸生成系统需要渲染视觉的、情感的表情。情感说话人脸生成的方法之一是从语音话语中估计表达的情感，然后将其呈现在生成的说话人脸中，但受语音情感识别准确性的限制，在视觉呈现中不能独立控制情感表达。文献[59]提出了一种不同的方法来解决这个问题，该方法忽略了语音音频中表达的情感，并将说话人脸的生成设定为一个独立的情感变量，提供了对视觉情感表达的直接和更灵活的控制，实现了更个性化的娱乐、教育和个人协助。它还为行为心理学家提供了一个强大的工具，可以进行以前不可能进行的与情绪相关的实验。例如，人们可以独立地研究人类如何通过音频和视觉方式操纵这些情绪来回应对话伙伴的情绪表达。其由一个神经网络组成，该神经网络可以从无条件情绪的言语中生成情绪化的说话人脸，并以语音、参考面部图像和分类情绪条件为输入，生成与带有情绪表达的输入语音同步的说话人脸。

文献[60]提出的语音操作的角色动画（voice operated character animation，VOCA）模型，使用任何语音信号，即使是非英语的语音，都可以逼真地生成各种各样的肖像的动画。通过训练，该模型学会了各种现实的说话风格，还提供了动画控制，以改变动画过程中的说话风格、身份依赖的面部形状和姿势。文献[60]指出，VOCA模型是唯一现实风格的适用于看不见的没有重新定位的对象目标的三维面部动画模型，它适用于游戏内视频，虚拟现实替身或任何事先不知道说话人、语音或语言的场景。

最近，跨模态的数据生成越来越受欢迎。有许多方法结合了音频和视觉，任务包括从视频中创建语音或从音频/语音中创建图像。语音驱动的人脸图像生成的研究主要采用非端到端的策略，并且利用先验数据知识。通常，经过专业开发以反映人类语言的手工特征被用于对语音进行编码，面部或嘴唇的点表示已被用于视觉部分。文献[61]专注于从音频/讲话中产生视觉效果，所提出的网络是完全端到端的训练，只使用未经处理的语音来产生图像像素，作者想要在未处理的语音波形上加一个GAN作为条件。该模型由3个端到端训练的模块组成：语音编码器、生成器网络和鉴别器网络。

在计算机视觉领域，从音频中创建嘴唇运动或从电影中合成移动的面部一直是一个有趣的任务。在大多数音频合成活动中，训练、建模或采样需要大量目标项目的视频素材，无法将随机的人脸图像与语音数据进行匹配。此外，如果没有关于面部和嘴唇运动的精确信息，现有方法无法产生良好的效果，这是一个典型的问题。文献[62]提出了解离的音视频系统（disentangled

audio-visual system，DAVS），它是一个端到端的用于生成说话人脸的可训练的深度网络。DAVS通过对抗学习分解出身份和语义信息，首先学习一个联合空间，将脸部序列和声音嵌入在一起。接下来强制执行唇读和语音识别的对齐，然后通过对抗学习分离，组合生成新的序列。

文献[63]解决了从源身份的音频/讲话和目标身份的简短视频（大约10 s长）中产生高质量谈话面部视频的问题，除理解从音频语音到嘴唇运动和面部表情（即头部姿势）的转换之外，此模型的说话人脸生成还考虑了目标身份独特的讲话行为。模型通过使用3D面部动画，以定制的头部姿势为核心，填补逼真的谈话脸部视频和视听驱动的头部姿势学习之间的空白。该模型的框架可以在以下两个阶段进行解释。第一阶段在传入音频、平均头部姿态和面部表情之间建立通用映射；然后，以输入视频为教学工具，对3D人脸进行重构，并调整泛型映射，学习独特的说话行为，以获取定制的头部姿势和3D面部动画。第二阶段使用输入视频中的纹理和光照数据将3D人脸动画渲染到视频帧中；然后，使用一个全新的内存增强GAN模块，将这些生成的帧调整成真实的帧。

大多数关于说话人脸生成的研究集中在通过唇形动作与输入音频同步来获取显式属性。音频转换和视觉生成网络将音频分解为与主题相关的信息和与语音相关的信息，以生成清晰的唇形[64]。此外，它将音频传输到面部地标，并根据地标生成视频帧。然而，最近只有少数研究关注了头部姿势的隐式属性与输入音频之间的关系。为了预测每个输入帧的变换矩阵，使用多层感知器作为头姿学习器。但是，下列情况仍不清楚：①如何用眨眼来模拟头部姿势这样的隐式属性；②外显属性和内隐属性是如何相互影响的。

文献[64]提出了一种FACIAL框架来解决上述问题。与以往使用单头姿态学习器预测内隐属性的研究不同，该框架采用对抗学习方法联合学习内隐属性和外显属性，它协同嵌入了所有属性，包括眨眼、头部姿势、表情、身份、纹理和光照的AU，从而可以在同一框架下建模它们对说话人脸生成的潜在交互。

文献[65]建议使用两个分析特征：wav2vec 2.0和Yingram。为了在没有任何文本信息的情况下保留语言信息，其利用了在53种语言上训练的wav2vec 2.0。此外，还提出了一个名为Yingram的新特征，其可比基频f_0更有效地表示和控制音高信息。虽然基频f_0主要用于描述基音数据，但当信号中存在次谐波时，f_0有时定义不清，其用一个可控制的抽象特性来解决上述问题。著名的Yin算法极大地启发了作者提出Yngram特征。上述分析特征有足够的数据来重构初始语音信号，结果发现，分析特征中的细节具有共同的知识，如音色和音高。文献[65]还提出了一种信息摄动方法，因为特定特征可以保持其预期目的的特定细节。这个想法是干扰所有的数据，从而训练神经网络不会从特征中提取不需要的属性，使该模型不再受到重建质量和特征解纠缠之间不可避免的权衡。

GAN相关的数字说话人脸的优势和劣势见**表1**。

2.3 使用Diffusion的数字说话人脸生成技术

Diffused Heads[70]首先使用扩散模型，只需要一个身份图像和音频序列就可以生成一个现实的会说话的人类头部的视频，包括产生头部运动、面部表情（如眨眼）

表1 GAN相关的数字说话人脸

模型	优势	缺点
RankGAN[34]	基于最大边际排序的渐进式；训练在后期阶段更加稳定，提高了GAN的收敛速度，增强了GAN的性能	在不增加容量的情况下提高性能
X2Face[55]	对看不见的姿势、表情、身份有鲁棒性；没有假设给出；它可以用作视频编辑设备	生成的图像质量不高
Mesoscopic Facial Geometry[35]	增强了精细细节的；计算纹理贴图和基础网格；产生一个非常详细的网格	需要条件输入，不是一个生成式3D人脸
High-fidelity Facial Reflectance[36]	输出详细的几何形状和高质量的反射率	需要条件输入，不是一个生成式3D人脸
Decoupled 3D face[38]	训练对齐的面部纹理，并结合线性3DMM生成逼真的合成数据	依赖于3DMM，而且形状基于情感
GANFit[3]	面部纹理的生成器在UV中训练；首次使用GAN进行模型拟合，得到高质量的面部纹理重建	面部纹理和形状无法重建高频法线
Synthesizing Facial Photometries[39]	学习人类身份和3D面部形状变化；二进制掩码提高效率	不模拟非线性，只关注面部表情
Unsupervised Face Normalization[66]	把3DMM拟合到图片中；GAN被用来完成UV映射中的缺失部分	生成的结果似乎分辨率很低
3DFaceGAN[40]	多标签数据；以2D人脸的UV映射、生成器检索面部特征为输入，生成接近实际目标的特征	只输入和输出高质量的3D数据
Digital Twin[43]	预测3D人脸模型的顶点坐标；使用监督学习技术估计高质量的3D人脸；3D人脸数据到3D人脸UV纹理数据映射	牙齿、眼睛、舌头和它们的模型架在比较之前被忽略
TB-GAN[67]	对GAN的结构和损失函数兼容；PG-GAN是根据WGAN-GP来训练的	不能正确地模拟人体皮肤和表面
AvatarMe[47]	从具有随机光照条件、姿势和遮挡的野生图像中重建高分辨率人脸	重建的反照率和法线似乎模糊不清
3D Face Generation via DCGAN[49]	既不需要预先计算的人脸对齐，也不需要显式的人脸特征提取；对3D人脸有效，因为它们的几何亲和性很高	只处理采样3D脸型，但不处理纹理
StyleGAN3[50]	无别名生成器包含关于给定训练数据性质的隐式假设	训练数据只剩下黑白像素
BlendGAN[68]	在统一模型中适合任意样式，同时避免逐个准备	训练数据只剩下黑白像素
MOST-GAN[69]	模拟人脸的物理属性，如3D形状、反照率、姿势和光照	改变光照条件对头发的影响有限
WAV2PIX	简短的语音片段；一种高效的跨模态技术生成人脸并合成可信表征，准确率为90.25%	该模型对图像的大小、语音块的长度以及最重要的训练数据的质量很敏感

的幻觉，并保留给定的背景。Diff2Lip[71]同样能够在保持这些特征的同时进行唇同步。

Difftalk[72]将参考脸图像和标记点作为补充条件来指导脸部身份和头部姿势的建模，从而实现身份信息的泛化，并且能够优雅地进行高分辨率合成，提高合成质量。DiffTalker[73]解决了与直接将扩散模型应用于音频控制相关的问题。DiffTalker由两个代理网络组成：基于变压器的几何精度地标补全网络和基于扩散的纹理细节人脸生成网络。地标在建立音频和图像域之间的无缝连接方面发挥着关键作用，促进了来自预训练扩散模型的知识的整合。

文献[74]的目标是，给定单人说话的视频，重新同步该人的嘴唇和下巴运动，以响应单独的听觉语音记录，而不依赖于中间结构表征，如面部标志或3D面部模型。

该方法还解决了多说话者情况下身份一致性的问题。

为了解决手工制作的中间表示不足的问题并精确描述面部运动，DAE-Talker[75]法利用从扩散自编码器（diffusion auto-encode，DAE）获得的数据驱动的潜在表示。DAE包含一个将图像编码为潜在向量的图像编码器和一个根据它重建图像的DDIM（denoising diffusion implicit model）图像解码器。Du等[75]在说话的人脸视频帧上训练DAE，然后提取其潜在表征作为基于Conformer的语音潜在模型的训练目标。这使DAE-Talker能够合成完整的视频帧，并产生与讲话内容一致的自然头部运动，而不是依赖于模板视频中预先确定的头部姿势。

FaceTalk[76]被用于从输入音频信号中合成高保真的人类说话头部3D运动序列。FaceTalk提出了一种新的潜在扩散模型，在神经参数头部模型的表达空间中操作，以合成音频驱动的真实头部序列。在缺乏具有与音频对应的NPHM（learning neural parametric head model）表达式的数据集的情况下，FaceTalk对这些对应进行优化，以生成适合人们说话的音频-视频记录的临时优化的NPHM表达式数据集。

EmoTalker[77]修改了去噪过程，以确保在推理过程中保留原始肖像的身份。为了提高文本输入的情感理解能力，引入了情感强度块来分析来自提示的细粒度情感和强度。

DiT-Head[78]基于扩散变压器，以音频为条件驱动扩散模型的去噪过程。DiT-Head方法具有可扩展性，可以推广到多个身份，同时生成高质量的结果。

文献[79]提出了一个带有属性引导扩散模型（face animation framework with an attribute-guided diffusion model，FADM）的人脸动画框架。为了解决扩散模型不可控的合成效应的问题，设计了一种属性导向条件网络（attention-driven graph clustering network，AGCN），自适应地结合粗动画特征和三维人脸重建结果，将外观和运动条件纳入扩散过程。这些特定的设计有助于FADM纠正不自然的工件和扭曲，并通过精确的动画属性迭代扩散细化，丰富高保真的面部细节。FADM可以灵活有效地改进现有的动画视频。

在文献[80]中，首先在文本提示符中表示情感，这可以继承CLIP（contrastive language-image pre-training）的丰富语义，从而实现灵活和广义的情感控制。作者进一步将这些任务重组为面向目标的纹理转移，并采用扩散模型。更具体地说，被提出的纹理几何感知扩散模型将复杂的传输问题分解为多条件去噪过程，其中，基于纹理注意力的模块精确地建模源条件和目标条件中包含的外观和几何线索之间的对应关系，并结合额外的隐含信息，以实现高保真的说话人脸生成。此外，此模型可以优雅地完成脸部交换。

DreamTalk[81]由3个关键部分组成：一个去噪网络、一个风格感知唇形专家和一个风格预测器。基于扩散的去噪网络能够在不同的表情中一致地合成高质量的音频驱动的面部运动。为了提高嘴唇动作的表现力和准确性，DreamTalk引入了一个风格感知嘴唇专家，可以指导口型同步，同时注意说话风格。为了消除对表情参考视频或文本的需求，使用了一个额外的基于扩散的风格预测器来直接从音频中预测目标表情，通过这种方式，DreamTalk可以利用强大的扩散模型来有效地生成富有表现力的面孔，并减少对昂贵的风格参考的依赖。

DREAM-Talk[82]是一个基于两阶段

扩散的音频驱动框架，专门用于将生成的不同的表情和精确的口型同步。在第一阶段，DREAM-Talk提出了EmoDiff，这是一个新颖的扩散模块，可以根据音频和参考的情感风格生成各种高度动态的情感表情和头部姿势。考虑到嘴唇运动和音频之间的强相关性，笔者使用音频特征和情感风格来改进动态，提高口型同步精度。为此，DREAM-Talk部署了一个视频到视频的渲染模块，将表情和嘴唇动作从代理3D化身转移到任意肖像。

Instruct-NeuralTalker[83]可以利用人类指令对说话辐射场进行编辑，从而实现个性化的说话人脸生成。该框架首先构建一个有效的说话辐射场，然后应用最新的条件扩散模型，在优化过程中使用人类指令编辑图像，以确保与编辑过程中的音频嘴唇同步。为了解决过度平滑的问题，该方法还引入了一个轻量级的细化网络，以补充图像细节，并在最终渲染图像中实现可控制的细节生成。

DiffPoseTalk[84]结合了从短参考视频中提取风格嵌入的风格编码器，在推理过程中，DiffPoseTalk采用基于语音和风格的无分类器引导和指导生成过程，文献[84]将其扩展到生成头部姿势，从而增强用户感知。

FaceDiffuser[85]是一个端到端的非确定性神经网络架构，用于语音驱动的3D面部动画合成。FaceDiffuser产生逼真和多样化的动画序列，并且可被推广到基于时间3D顶点的网格动画数据集和基于时间混合形状的数据集。

3DiFACE[86]提出了一个用于3D面部运动的轻量级音频条件扩散模型。这种扩散模型可以在一个小的3D运动数据集上训练，保持富有表现力的嘴唇运动输出。此外，它还可以针对特定的主题进行微调，只需要一个人的短视频。

2.4 使用神经辐射场的数字说话人脸生成技术

传统的说话人脸生成模型只能生成具有固定头部姿势的音频同步嘴唇运动。最近的一些作品考虑个性化属性来解决这个问题。然而，这些方法使用确定性模型来生成个性化信息，结果缺乏多样性，导致模式重复。神经辐射场最近展现出了以无限分辨率渲染高质量的说话面部图像的能力。然而，这些作品忽略了个性化的面部属性，无法准确地将音频与嘴唇运动同步。神经网络在神经场景表示中用于表示场景的形式和外观。一个可以在空间不同点采样的神经网络，神经场景表示网络（neural scene representation network，SRN）表示物体的几何形状和外观。对于大脑渲染和重建的挑战，神经辐射场引起了人们的极大兴趣。体积射线采样结果可以从3D对象的形式和外观的隐式表示中创建。使用野外训练数据、外观插值、可变形神经辐射场来表示非刚性运动物体，神经辐射场（neural radiance field，NeRF）在没有预先计算的相机参数的情况下进行优化，随后的其他研究进一步推进了这一概念[2]。

3 与数字说话人脸相关的数据集和评价指标

与视听数据集相关的问题实质上影响了说话人脸生成的研究进展。因为目前大多数深度学习算法是数据驱动的，数据集的重要性不言而喻。另外，由于隐私问题和昂贵的人工成本，现有的视听数据集受到小规模和不充分的注释的影响。一个前瞻性的研究方向是基于无标记视听数据的跨模态自监督视觉语音学习。尽管如此，数

据集规模受到限制的问题仍然没有得到解决。大规模视听数据集为人脸对话视频制作方面的最新成果做出了重大贡献，可以使用越来越丰富的数据集来训练强大的说话面部图像，这些数据集可以捕获照明环境、身份、视频质量和句子中的各种视觉内容，且数据集的音频和注释特征为检查和对比各种方法的有效性提供了完整的方法。最近发布的与语音相关的视听数据集的特征见**表2**，其范围从实验室控制到野外环境数据。

许多3D面部数据集已被公开，目的是分析静态或动态的面部表情。这些数据集中只有少数捕捉到了由语言引起的面部动态，大多数集中于情感表情。后来发布的4DFAB数据集包括180个受试者的4D捕捉，但网格质量较差，每个受试者只有9个词的发言。以下是3D的高质量面部数据集。

● B3D(AC)2数据集[60]包含了40个英语口语句子片段的大量音频4D扫描配对。有些句子由多个受试者说出，有些则只由一个受试者说出，以便对各种句子和受试者进行训练。由于原始B3D(AC)2扫描中存在明显的伪影，注册模板只覆盖了面部，省略了与语言相关的颈部动作，这可能导致微妙的面部动作丢失。

● 众包情感多模态演员数据集（crowd-sourced emotional multimodal actors dataset，CREMA-D）[87]包含91位演员（48位男性演员和43位女性演员），表达6种分类情感的视频片段：愤怒、厌恶、恐惧、快乐、中性和悲伤。演员的年龄范围为20岁至74岁。每个视频片段展示了一位演员用其中一种情感类别说出12个句子中的一个。提供的视频分辨率为480×360，采样率为每秒30帧。音频的采样率为44.1 kHz。

● 油管博主数据集[88]由两部分组成：一部分是自动生成的完整噪声子集；另一部分是手动策划的干净子集，以获取高质量数据。总共收集了168 796 s的语音，以及与过去几年活跃的62位油管博主相

表 2 与语音相关的视听数据集的特征

数据集	时间长度/h	主题/个	词库	情绪	收集环境
LRS3-TED	438	5 000+	N/A	X	野外
MELD	13.7	407	17 000	+	野外
GRID	27.5	33	51	X	实验室
TCD-TIMIT	11.1	62	N/A	X	实验室
MODALITY	31	35	182	X	实验室
LRW	173	1 000+	500	X	野外
CREMA-D	11.1	91	N/A	+	实验室
RAVDESS	7	24	8	+	实验室
MSP-IMPROV	18	12	N/A	+	实验室
VoxCeleb2	24 000	61 000	N/A	X	野外
LRS2-BBC	224.5	500+	59 000	X	野外
Faceforensics++	5.7	1k	N/A	X	野外
ObamaSet	14	1	N/A	X	野外
VoxCeleb1	352	1.2	N/A	X	野外
HDTF	15.8	300+	N/A	X	野外

关的视频帧和裁剪后的面部图像。该数据集性别均衡，并经过手动清理，保留了42 199张面部图像，每张都与1 s的语音片段相关联。

● TCD-TIMID数据集[89]包含62位演讲者发表的6 913个音素丰富句子的高质量音频和视频，这些句子在视频中没有可见的头部变化。视频是从正面和30度的角度录制的。

● GRID（ground re-ldentification）数据集[90]中包括33位面对摄像头的演讲者，每人说出1 000个短语。它包含了从有限词典中随机选择的6个单词，总共51个单词。句子以中性情感说出，没有任何明显的头部动作。

● MODALITY数据集[91]使用飞行时间（time-of-flight）相机提供额外的面部深度数据分析。使用SoftKinetic DepthSense 325相机模型，以每秒60帧的速度提供深度数据，空间分辨率为320×240像素。除此之外，该相机还可以从立体RGB相机录像中检索3D数据。

● LRW（lip reading in the wild）数据集[92]提供了500个不同的单词，由不同的演讲者在野外说出。这个数据集涵盖了各种头部姿势，一些视频中包括一个演讲者直接对着摄像头说话。相比之下，其他视频包含了演讲者在小组辩论中互相看着对方，指出一些视频中有严重的头部姿势。由于LRW数据集是在现实世界中收集的，并附有真实标签，所以视频较短。

除说话人脸生成领域的困难之外，有效地评估生成方法也是一个问题。而针对这一问题，笔者的方向在于：①目标个人的声音应该与合成视频的说话者的声音相匹配；②制作的视频质量要高；③说话时应该眨眼；④创建的说话者的嘴型应该与音频相似等。

某些技术在处理不正确或丢失的说话者身份方面存在问题，如目标说话者和生成的说话者看起来并非同一人。因此，在创建说话面孔的任务中，也使用了一些评估身份保留的措施。通常使用预训练的人脸识别模型作为身份特征提取器，通过测量特征之间的距离来量化身份保持。例如，平均内容距离（average content distance，ACD）是通过测量参考身份图像和预测图像的FaceNet特征之间的相似度来计算的，余弦相似度（cosine similarity，CSIM）用于测量ArcFace的嵌入向量之间的身份不匹配度[93]。

在人脸生成领域，经常使用图像质量评估标准。通过均方误差定义的峰值信噪比（peak signal-to-noise ratio，PSNR）可以表示两幅图片之间的像素级差异。然而，人类感知与PSNR之间的差异仍然较大。通过结构相似度（structural similarity，SSIM）测量两幅图像之间的亮度、对比度和结构的相似性，引入了IS（inception score）来评估生成模型的多样性，通过比较由训练过的Inception-v3模型产生的两个特征的平均值和标准差，确定了FID。然而，为了评估这些技术，需要参考照片。使用非参考图像评估指标-累积概率模糊检测（cumulative probability blur detection，CPBD）来评估图像的清晰度。相比之下，频域模糊度量（frequency domain blurriness measure，FDBM）基于图像频谱评估频域模糊度。

同样重要的是眨眼的真实性。生成的视频中的平均眨眼时间和频率应该与真实人类的眨眼时间和频率相似。研究者计算了平均持续时间和频率，以评估眨眼的自然程度。评估说话面部合成自然度的其他关键指标是音频唇部同步。口部关键点距离（landmark distance，LMD）用于测量生成的口形，是生成图像与现实世界参考图像之间的差距。唇读练习学习了从面

部照片到相应文本的映射，训练有素的唇读模型可以确定单词错误率（word error rate，WER）。Syncnet模型确定并评估音频视觉同步，更好的唇部同步表现为降低的AV偏移和更高的AV信心。从人类感知的角度来看，唇读相似度距离（lip-reading similarity distance，LRSD）是一个唇部同步评估指标，它基于一个唇读模型，通过比较生成视频片段与实际视频片段的特征来测量音频-视觉是否同步。

数字说话人脸相关的评价指标见**表3**。

4 结束语

数字说话人脸生成技术是计算机视觉中一个重要而富有挑战性的课题，受到了广泛的关注。由于生成模型的研究取得了显著的进展，深度学习方法（如CNN、GAN和NeRF）技术领域的说话人脸生成进步很快。本文综述了上述生成模型的研究现状及其在说话人脸生成中的实现，介绍了几个观点和研究方向，让读者对这些主题有更广泛的了解；并对说话人脸生成和动画任务及其评价指标进行了全面的综述，重点介绍了近年来的研究成果，为自然说话人脸生成提供了定义良好的模型。此外，本文还列出了现有的数据集，对可用方法进行了比较和分析，并使用先进的阈值在相同条件下对模型进行了基准测试。未来的研究目标包括实时语音驱动的3D对话人脸生成，包括AR、VR、远程会议、计算机游戏和虚拟穿戴等。

参考文献：

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

[2] 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 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2021: 5764-5774.

[3] GECER B, PLOUMPIS S, KOTSIA I, et al. GANFIT: generative adversarial network fitting for high fidelity 3D face reconstruction[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 1155-1164.

[4] HU G, CHAN C H, KITTLER J, et al. Resolution-aware 3D morphable model[C]//Proceedings of BMVC. Guildford: Springer, 2012: 1-10.

[5] YIN X, YU X, SOHN K, et al. Towards large-pose face frontalization in the wild[C]//Proceedings of the 2017 IEEE International Conference on Computer

表 3 说话人脸相关的评价指标

项目	评价指标	
	更高好	更低好
身份保存	CSIM, ID, SSIM, PSNR	FID
图片质量保存	FDBM	ACD
音唇同步	音视频置信度	LRSD, AV(Offset), WER, LMD
眨眼	期间的频率	

Vision (ICCV). Piscataway: IEEE Press, 2017: 4010–4019.

[6] DAI H, PEARS N, SMITH W, et al. Statistical modeling of craniofacial shape and texture[J]. International Journal of Computer Vision, 2020, 128(2): 547–571.

[7] SZABÓ A, MEISHVILI G, FAVARO P. Unsupervised generative 3D shape learning from natural images[EB]. arXiv preprint, 2019, arXiv: 1910.00287.

[8] WU Z J, WANG X, LIN D, et al. SAGNet: structure–aware generative network for 3D–shape modeling[J]. ACM Transactions on Graphics, 2019, 38(4): 91.

[9] DEPRELLE T, GROUEIX T, FISHER M, et al. Learning elementary structures for 3D shape generation and matching[EB]. arXiv preprint, 2019, arXiv: 1908.04725.

[10] ALGADHY R, GOTOH Y, MADDOCK S. 3D visual speech animation using 2D videos[C]//Proceedings of the ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2019: 2367–2371.

[11] ZHONG W Z, FANG C W, CAI Y Q, et al. Identity–preserving talking face generation with landmark and appearance priors[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 9729–9738.

[12] BOOTH J, ROUSSOS A, ZAFEIRIOU S, et al. A 3D morphable model learnt from 10, 000 faces[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 5543–5552.

[13] TRAN L, LIU X M, TRAN L, et al. On learning 3D face morphable model from In–the–wild images[EB]. arXiv preprint, 2018, arXiv: 1808.09560.

[14] TRAN L, LIU X M. On learning 3D face morphable model from In–the–wild images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(1): 157–171.

[15] DAI H, PEARS N, SMITH W, et al. A 3D morphable model of craniofacial shape and texture variation[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 3104–3112.

[16] TRUONG K T. Video–based face recognition using shape and texture information in 3d morphable model[J]. JP Journal of Heat and Mass Transfer, 2018, SV2018(1): 119–124.

[17] DENG J K, CHENG S Y, XUE N N, et al. UV–GAN: adversarial facial UV map completion for pose–invariant face recognition[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7093–7102.

[18] DHANWADA K R, DICKENS M, NEADES R, et al. Differential effects of UV–B and UV–C components of solar radiation on MAP kinase signal transduction pathways in epidermal keratinocytes[J]. Oncogene, 1995, 11(10): 1947–1953.

[19] BOOTH J, ZAFEIRIOU S. Optimal UV spaces for facial morphable model construction[C]//Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE Press, 2014: 4672–4676.

[20] WU F Z, BAO L C, CHEN Y J, et al. MVF–net: multi–view 3D face morphable model regression[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 959–968.

[21] BOOTH J, ROUSSOS A, PONNIAH A, et al. Large scale 3D morphable models[J]. International Journal of Computer Vision, 2018, 126(2): 233–254.

[22] WU F Z, LI S N, ZHAO T H, et al.

Cascaded regression using landmark displacement for 3D face reconstruction[J]. Pattern Recognition Letters, 2019, 125: 766-772.

[23] TRAN L, LIU F, LIU X M. Towards high-fidelity nonlinear 3D face morphable model[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 1126-1135.

[24] HU G S, YAN F, CHAN C H, et al. Face recognition using a unified 3D morphable model[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2016: 73-89.

[25] CATE H, DALVI F, HUSSAIN Z. DeepFace: face generation using deep learning[EB]. arXiv preprint, 2017, arXiv: 1701.01876.

[26] WANG W S, XI J Q, HEDRICK J K. A learning-based personalized driver model using bounded generalized Gaussian mixture models[J]. IEEE Transactions on Vehicular Technology, 2019, 68(12): 11679-11690.

[27] ZHU X P, BHUIYAN A, MEKHALFI M L, et al. Exploiting Gaussian mixture importance for person re-identification[C]//Proceedings of the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Piscataway: IEEE Press, 2017: 1-6.

[28] HAMDI A, GHANEM B. IAN: combining generative adversarial networks for imaginative face generation[EB]. arXiv preprint, 2019, arXiv: 1904.07916.

[29] PUMAROLA A, AGUDO A, MARTINEZ A M, et al. GANimation: anatomically-aware facial animation from a single image[J]. Computer Vision - ECCV: European Conference on Computer Vision: Proceedings European Conference on Computer Vision, 2018, 11214: 835-851.

[30] ZHOU Y Q, SHI B E. Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder[C]// Proceedings of the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). Piscataway: IEEE Press, 2017: 370-376.

[31] 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 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 8652-8661.

[32] DI X, SINDAGI V A, PATEL V M. GP-GAN: gender preserving GAN for synthesizing faces from landmarks[C]// Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR). Piscataway: IEEE Press, 2018: 1079-1084.

[33] PRAJWAL K R, MUKHOPADHYAY R, PHILIP J, et al. Towards automatic face-to-face translation[EB]. arXiv preprint, 2020, arXiv: 2003.00418.

[34] JUEFEI-XU F, DEY R, BODDETI V N, et al. RankGAN: a maximum margin ranking GAN for generating faces[C]//Proceedings of Asian Conference on Computer Vision. Cham: Springer, 2019: 3-18.

[35] HUYNH L, CHEN W K, SAITO S, et al. Mesoscopic facial geometry inference using deep neural networks[C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8407-8416.

[36] YAMAGUCHI S, SAITO S, NAGANO K, et al. High-fidelity facial reflectance and geometry inference from an unconstrained image[J]. ACM Transactions on Graphics, 2018, 37(4): 162.

[37] ZHANG H, GOODFELLOW I, METAXAS

D, et al. Self-attention generative adversarial networks[EB]. arXiv preprint, 2018: arXiv: 1805.08318.

[38] ABREVAYA V F, BOUKHAYMA A, WUHRER S, et al. A decoupled 3D facial shape model by adversarial training[C]// Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 9418-9427.

[39] SHAMAI G, SLOSSBERG R, KIMMEL R. Synthesizing facial photometries and corresponding geometries using generative adversarial networks[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2019, 15(3s): 87.

[40] MOSCHOGLOU S, PLOUMPIS S, NICOLAOU M A, et al. 3DFaceGAN: adversarial nets for 3D face representation, generation, and translation[J]. International Journal of Computer Vision, 2020, 128(10): 2534-2551.

[41] 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 - SIGGRAPH '99. New York: ACM, 1999: 187-194.

[42] CAO C, WENG Y L, ZHOU S, et al. FaceWarehouse: a 3D facial expression database for visual computing[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(3): 413-425.

[43] WANG R Z, CHEN C F, PENG H, et al. Digital twin: acquiring high-fidelity 3D avatar from a single image[EB]. arXiv preprint, 2019, arXiv: 1912.03455.

[44] JIA R M, LI T, YUAN F. FF-GAN: feature fusion GAN for monocular depth estimation[M]. Pattern Recognition and Computer Vision. Cham: Springer, 2020: 167-179.

[45] HUANG R, ZHANG S, LI T Y, et al. Beyond face rotation: global and local perception GAN for photorealistic and identity preserving frontal view synthesis[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 2458-2467.

[46] WEN Y, RAJ B, SINGH R. Face reconstruction from voice using generative adversarial networks[J]. Advances in Neural Information Processing Systems, 2019, 32.

[47] LATTAS A, MOSCHOGLOU S, GECER B, et al. AvatarMe: realistically renderable 3D facial reconstruction "in-the-wild"[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle: IEEE, 2020: 760-769.

[48] GECER B, LATTAS A, PLOUMPIS S, et al. Synthesizing coupled 3D face modalities by trunk-branch generative adversarial networks[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2020: 415-433.

[49] LUO G L, ZHAO X, TONG Y, et al. Geometry sampling for 3D face generation via DCGAN[C]//Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE Press, 2020: 1-7.

[50] KARRAS T, AITTALA M, LAINE S, et al. Alias-free generative adversarial networks[EB]. arXiv preprint, 2021, arXiv: 2106.12423.

[51] KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks[C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 4396-4405.

[52] VAN DEN OORD A, KALCHBRENNER

N, KAVUKCUOGLU K. Pixel recurrent neural networks[EB]. arXiv preprint, 2016, arXiv: 1601.06759.

[53] VAN DEN OORD A, KALCHBRENNER N, VINYALS O, et al. Conditional image generation with PixelCNN decoders[EB]. arXiv preprint, 2016, arXiv: 1606.05328.

[54] JAMALUDIN A, CHUNG J S, ZISSERMAN A. You said that? : synthesising talking faces from audio[J]. International Journal of Computer Vision, 2019, 127(11): 1767–1779.

[55] WILES O, KOEPKE A S, ZISSERMAN A. X2Face: A network for controlling face generation using images, audio, and pose codes[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2018: 690–706.

[56] EPHRAT A, HALPERIN T, PELEG S. Improved speech reconstruction from silent video[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). Piscataway: IEEE Press, 2017: 455–462.

[57] CHUNG J S, SENIOR A, VINYALS O, et al. Lip reading sentences in the wild[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 3444–3453.

[58] PHAM H X, WANG Y T, PAVLOVIC V. Generative adversarial talking head: bringing portraits to life with a weakly supervised neural network[EB]. arXiv preprint, 2018, arXiv: 1803.07716.

[59] ESKIMEZ S E, ZHANG Y, DUAN Z Y. Speech driven talking face generation from a single image and an emotion condition[J]. IEEE Transactions on Multimedia, 2022, 24: 3480–3490.

[60] CUDEIRO D, BOLKART T, LAIDLAW C, et al. Capture, learning, and synthesis of 3D speaking styles[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 10093–10103.

[61] DUARTE A, ROLDAN F, TUBAU M, et al. Wav2Pix: speech-conditioned face generation using generative adversarial networks[C]//Proceedings of the ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2019: 8633–8637.

[62] ZHOU H, LIU Y, LIU Z W, et al. Talking face generation by adversarially disentangled audio-visual representation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 9299–9306.

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

[64] ZHANG C X, ZHAO Y F, HUANG Y F, et al. FACIAL: synthesizing dynamic talking face with implicit attribute learning[C]// Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2021: 3847–3856.

[65] CHOI H S, LEE J, KIM W, et al. Neural analysis and synthesis: reconstructing speech from self-supervised representations[EB]. arXiv preprint, 2021, arXiv: 2110.14513.

[66] QIAN Y C, DENG W H, HU J N. Unsupervised face normalization with extreme pose and expression in the wild[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 9843–9850.

[67] GECER B, LATTAS A, PLOUMPIS S, et al. Synthesizing coupled 3d face modalities by trunk-branch generative

adversarial networks[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 415–433.

[68] LIU M C, LI Q, QIN Z K, et al. BlendGAN: implicitly GAN blending for arbitrary stylized face generation[EB]. arXiv preprint, 2021, arXiv: 2110.11728.

[69] MEDIN S C, EGGER B, CHERIAN A, et al. MOST-GAN: 3D morphable StyleGAN for disentangled face image manipulation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(2): 1962–1971.

[70] STYPUŁKOWSKI M, VOUGIOUKAS K, HE S, et al. Diffused heads: diffusion models beat GANs on talking-face generation[C]// Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE Press, 2024: 5089–5098.

[71] MUKHOPADHYAY S, SURI S, GADDE R T, et al. Diff2Lip: audio conditioned diffusion models for lip-synchronization[C]//Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE Press, 2024: 5280–5290.

[72] SHEN S, ZHAO W L, MENG Z B, et al. DiffTalk: crafting diffusion models for generalized audio-driven portraits animation[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 1982–1991.

[73] QI Z, ZHANG X, CHENG N, et al. DiffTalker: co-driven audio-image diffusion for talking faces via intermediate landmarks[EB]. arXiv preprint, 2023, arXiv: 2309.07509.

[74] BIGIOI D, BASAK S, STYPUŁKOWSKI M, et al. Speech driven video editing via an audio-conditioned diffusion model[J]. Image and Vision Computing, 2024, 142: 104911.

[75] DU C P, CHEN Q, HE T Y, et al. DAE-talker: high fidelity speech-driven talking face generation with diffusion autoencoder[C]//Proceedings of the Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM, 2023: 4281–4289.

[76] ANEJA S, THIES J, DAI A, et al. Facetalk: audio-driven motion diffusion for neural parametric head models[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 21263–21273.

[77] ZHANG B Y, ZHANG X L, CHENG N, et al. EmoTalker: emotionally editable talking face generation via diffusion model[C]//Proceedings of the ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2024: 8276–8280.

[78] MIR A, ALONSO E, MONDRAGÓN E. DiT-head: high-resolution talking head synthesis using diffusion transformers[EB]. arXiv preprint, 2023, arXiv: 2312.06400.

[79] ZENG B H, LIU X H, GAO S C, et al. Face animation with an attribute-guided diffusion model[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE Press, 2023: 628–637.

[80] XU C, ZHU S, ZHU J, et al. Multimodal-driven talking face generation via a unified diffusion-based generator[EB]. arXiv preprint, 2023, arXiv: 2305.02594.

[81] MA Y F, ZHANG S W, WANG J Y, et al. DreamTalk: when expressive talking head generation meets diffusion probabilistic models[EB]. ArXiv preprint, 2023: arXiv: 2312.09767.

[82] ZHANG C, WANG C, ZHANG J, et al. DREAM-Talk: diffusion-based realistic emotional audio-driven method for single image talking face generation[EB]. arXiv preprint, 2023, arXiv: 2312.13578.

[83] SUN Y, HE R, TAN W, et al. Instruct-NeuralTalker: editing audio-driven talking radiance fields with instructions[EB]. arXiv preprint, 2023, arXiv: 2306.10813.

[84] SUN Z, LV T, YE S, et al. Diffposetalk: speech-driven stylistic 3d facial animation and head pose generation via diffusion models[EB]. arXiv preprint, 2023, arXiv: 2310.00434.

[85] STAN S, HAQUE K I, YUMAK Z. FaceDiffuser: speech-driven 3D facial animation synthesis using diffusion[C]// Proceedings of the ACM SIGGRAPH Conference on Motion, Interaction and Games. New York: ACM, 2023: 1-11.

[86] THAMBIRAJA B, ALIAKBARIAN S, COSKER D, et al. 3DiFACE: diffusion-based speech-driven 3D facial animation and editing[EB]. arXiv preprint, 2023, arXiv: 2312.00870.

[87] ESKIMEZ S E, MADDOX R K, XU C L, et al. End-to-end generation of talking faces from noisy speech[C]//Proceedings of the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2020: 1948-1952.

[88] MALIK S, RATHEE C, WANG T Y. YouTubers balancing the paradox of novelty and conformity[J]. Academy of Management Proceedings, 2020, 2020(1): 18544.

[89] MORRONE G, BERGAMASCHI S, PASA L, et al. Face landmark-based speaker-independent audio-visual speech enhancement in multi-talker environments[C]//Proceedings of the ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2019: 6900-6904.

[90] SONG J F, YANG Y X, SONG Y Z, et al. Generalizable person re-identification by domain-invariant mapping network[C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 719-728.

[91] PENG C L, WANG N N, LI J, et al. DLFace: deep local descriptor for cross-modality face recognition[J]. Pattern Recognition, 2019, 90: 161-171.

[92] MESBAH A, BERRAHOU A, HAMMOUCHI H, et al. Lip reading with Hahn convolutional neural networks[J]. Image and Vision Computing, 2019, 88: 76-83.

[93] WANG Y X, SONG L S, WU W, et al. Talking faces: audio-to-video face generation[M]. Handbook of Digital Face Manipulation and Detection. Cham: Springer, 2022: 163-188.

作者简介

张冰源(1999-),男,中国科学技术大学先进技术研究院硕士生,平安科技(深圳)有限公司实习生,主要研究方向为大语言模型、可解释性、Talking Face等。

张旭龙（1988- ），男，博士，平安科技（深圳）有限公司高级算法研究员，清华大学深圳研究院、中国科学技术大学先进技术研究院校外导师，IEEE、中国自动化学会、中国计算机学会会员，联邦数据与联邦智能专委会委员，主要研究方向为语音合成、语音转换、音乐信息检索、机器学习、深度学习方法在人工智能领域的应用。2023年入选上海市东方英才计划青年项目。

王健宗（1983- ），男，博士，平安科技（深圳）有限公司副总工程师，资深人工智能总监，联邦学习技术部总经理，智能金融前沿技术研究院院长。美国佛罗里达大学人工智能博士后，美国莱斯大学和华中科技大学联合培养博士，中国计算机学会资深会员，中国计算机学会大数据专家委员会委员，中国自动化学会联邦数据和联邦智能专业委员会副主任。主要研究方向为大模型、联邦学习和深度学习等。

程宁（1981- ），男，博士，平安科技（深圳）有限公司高级人工智能专家，主要研究方向为人工智能算法研究及其在语音处理和自然语言处理领域的应用。在大数据、机器学习、人工智能国际顶会或期刊上发表学术论文50余篇，发明专利申请100余项。

肖京（1972- ），男，美国卡耐基梅隆大学博士，IEEE Fellow，国家特聘专家。国家新一代普惠金融人工智能开放创新平台技术负责人、深圳市政协委员、深圳市决策咨询委员会委员，兼中国计算机学会深圳分部副主席、广东省人工智能与机器人学会副理事长、深圳市人工智能行业协会会长、深圳市人工智能学会副理事长， 清华大学、上海交通大学、同济大学等客座教授。长期从事人工智能与大数据分析挖掘相关领域研究，先后在爱普生美国研究院及美国微软公司担任高级研发管理职务，现任平安集团首席科学家，负责人工智能技术研发及其在金融、医疗、智慧城市等领域的应用，带领团队树立了多项传统行业智能化经营的标杆。已发表学术论文249篇，美国授权专利101项，中国发明专利155项，参与及承担国家级项目8项。凭借在技术创新及应用的杰出贡献，先后获得2018年中国专利奖、2019年吴文俊人工智能杰出贡献奖、2020年吴文俊人工智能科技进步一等奖、2020年上海市科技进步奖一等奖、2020年中国人工智能十大风云人物、2021年深圳市五一劳动奖章、2022年深圳市最美科技工作者等荣誉。

收稿日期： 2023-09-26
通信作者： 王健宗，jzwang@188.com
基金项目： 广东省重点领域研发计划“新一代人工智能”重大专项（No.2021B0101400003）
Foundation Item: The Key Research and Development Program of Guangdong Province (No.2021B0101400003)