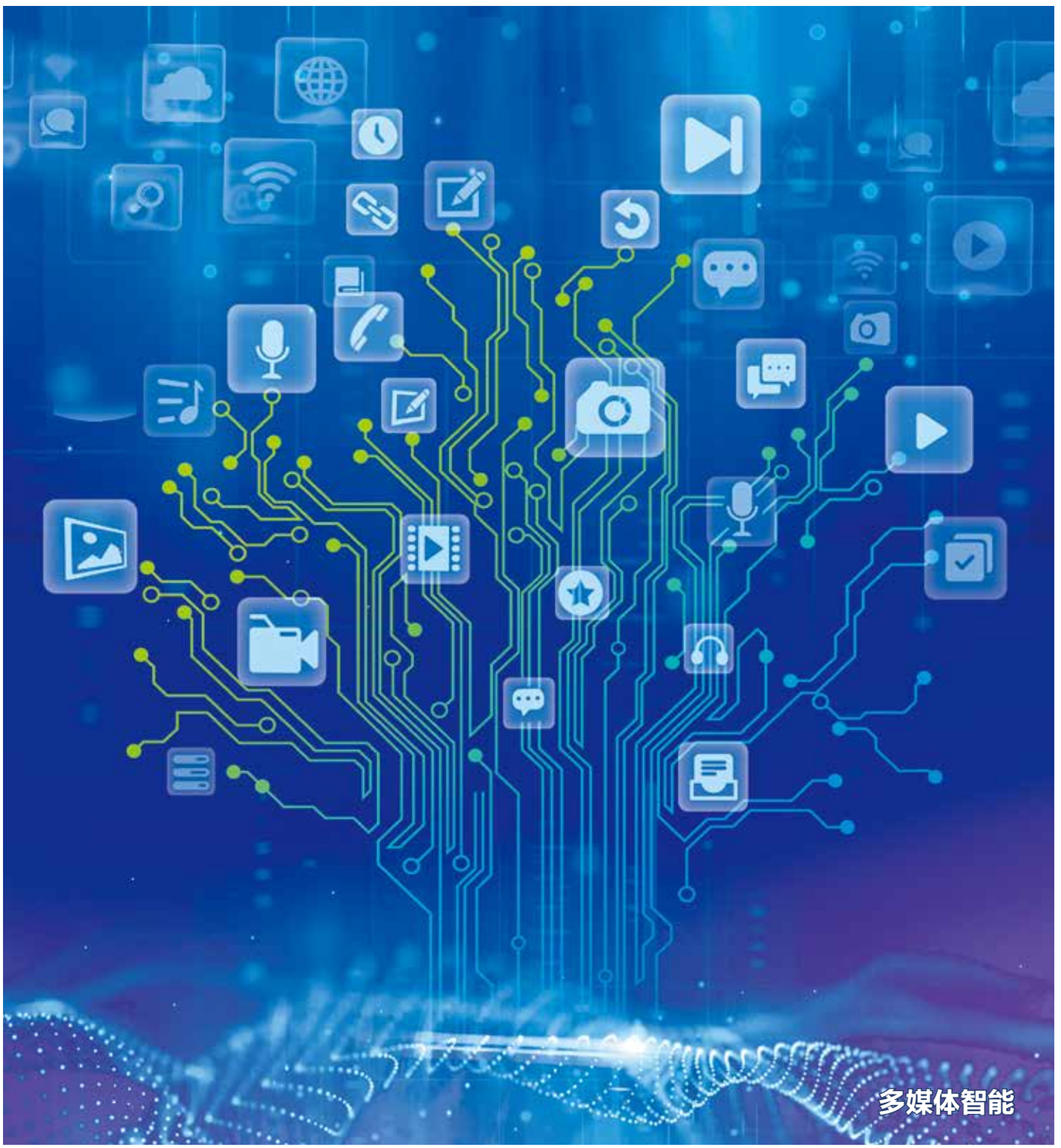
JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2022
09
VOL.27

ISSN1006-8961
CN11-3758/TB



中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第27卷第9期（总第317期）

2022年9月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@aircas.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

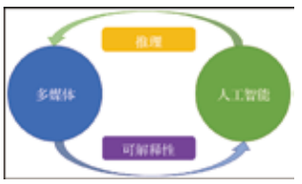
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2022年9月 第27卷 第9期（总第317期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

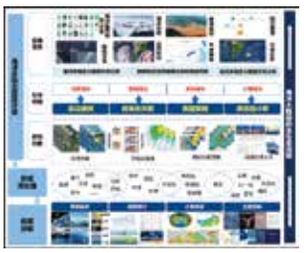
目次



多媒体智能：当多媒体遇到人工智能(第2551页)

学者观点

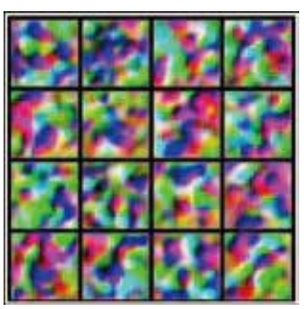
综述



面向海洋的多模态智能计算：挑战、进展和展望(第2589页)

多媒体智能安全

目标智能检测

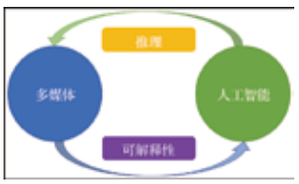


基于真实数据感知的模型功能窃取攻击(第2721页)

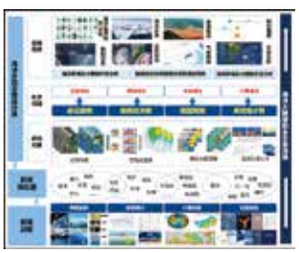
多媒体分析与理解

Volume 27, Number 9
Published September 16, 2022

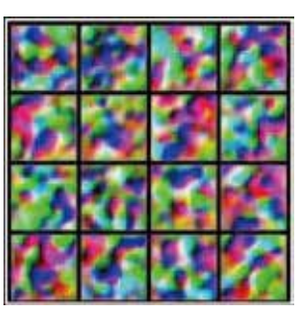
CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



Multimedia intelligence: the convergence of multimedia and artificial intelligence(P2551)



Marine oriented multimodal intelligent computing: challenges, progress and prospects(P2589)



Model functionality stealing attacks based on real data awareness(P2721)

中图法分类号:TP18　　**文献标识码**: A　　**文章编号**: 1006-8961(2022)09-2629-23

论文引用格式: Liu J, Chen P, Wang X, Fu X M, Dai J and Han J Z. 2022. Critical review of human face reenactment methods. Journal of Image and Graphics,27(09):2629-2651(刘锦,陈鹏,王茜,付晓蒙,戴娇,韩冀中. 2022. 人类面部重演方法综述. 中国图象图形学报,27(09):2629-2651)
[DOI:10.11834/jig.211243]

人类面部重演方法综述

刘锦[1,2],陈鹏[1,2],王茜[1*],付晓蒙[1,2],戴娇[1],韩冀中[1]

1. 中国科学院信息工程研究所,北京　100093; 2. 中国科学院大学网络空间安全学院,北京　100049

摘　要:　随着计算机视觉领域图像生成研究的发展,面部重演引起广泛关注,这项技术旨在根据源人脸图像的身份以及驱动信息提供的嘴型、表情和姿态等信息合成新的说话人图像或视频。面部重演具有十分广泛的应用,例如虚拟主播生成、线上授课、游戏形象定制、配音视频中的口型配准以及视频会议压缩等,该项技术发展时间较短,但是涌现了大量研究。然而目前国内外几乎没有重点关注面部重演的综述,面部重演的研究概述只是在深度伪造检测综述中以深度伪造的内容出现。鉴于此,本文对面部重演领域的发展进行梳理和总结。本文从面部重演模型入手,对面部重演存在的问题、模型的分类以及驱动人脸特征表达进行阐述,列举并介绍了训练面部重演模型常用的数据集及评估模型的评价指标,对面部重演近年研究工作进行归纳、分析与比较,最后对面部重演的演化趋势、当前挑战、未来发展方向、危害及应对策略进行了总结和展望。

关键词:人工智能(AI);计算机视觉;深度学习;生成对抗网络(GAN);深度伪造;面部重演

Critical review of human face reenactment methods

Liu Jin[1,2], Chen Peng[1,2], Wang Xi[1*], Fu Xiaomeng[1,2], Dai Jiao[1], Han Jizhong[1]

1. Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China;

2. School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: Current image and video data have been increasing dramatically in terms of huge artificial intelligence (AI)-generated contents. The derived face reenactment has been developing based on generated facial images or videos. Given source face information and driving motion information, face reenactment aims to generate a reenacted face or corresponding reenacted face video of driving motion information in related to the animation of expression, mouth shape, eye gazing and pose while preserving the identity information of the source face. Face reenactment methods can generate a variety of multiple feature-based and motion-based face videos, which are widely used with less constraints and becomes a research focus in the field of face generation. However, almost no reviews are specially written for the aspect of face reenactment. In view of this, we carry out the critical review of the development of face reenactment beyond DeepFake detection contexts. Our review is focused on the nine perspectives as following: 1) the universal process of face reenactment model; 2) facial information representation; 3) key challenges and barriers; 4) the classification of related methods; 5) introduction of various face reenactment methods; 6) evaluation metrics; 7) commonly used datasets; 8) practical applications; and 9) conclusion and future prospect. The identity information and background information is extracted from source faces while motion

收稿日期:2022-01-20;**修回日期**:2022-05-25;**预印本日期**:2022-06-02

*通信作者:王茜　wangxi1@iie.ac.cn

基金项目:科技创新2030-"新一代人工智能"重大项目(2020AAA0140000);国家自然科学基金项目(61702502)

Supported by: National Key R&D Program of China (2020AAA0140000); National Natural Science Foundation of China (61702502)

features are extracted from driving information, which are combined to generate the reenacted faces. Generally, latent codes, 3D morphable face models (3DMM) coefficients, facial landmarks and facial action units are all served as motion features. Besides, there exist several challenges and problems which are always focused in related research. The identity mismatch problem means the inability of face reenactment model to preserve the identity of source faces. The issue of temporal or background inconsistency indicates that the generated face videos are related to the cross-framing jitter or obvious artifacts between the facial contour and the background. The constraints of identity are originated from the model design and training procedure, which can merely reenact the specific person seen in the training data. As for the category of face reenactment methods, image-driven methods and cross-modality-driven methods are involved according to the modality of driving information. Based on the difference of driving information representation, image-driven methods can be divided into four categories. The driving information representation includes facial landmarks, 3DMM, motion field prediction and feature decoupling. The subclasses of identity restriction (yes/no issue) can be melted into the landmark-based and 3DMM-based methods further in terms of whether the model could generate unseen subjects or not. Our demonstration of each category, corresponding model flowchart and following improvement work will be illustrated in detail. As for the cross-modality driven methods, the text and audio related methods are introduced, which are ill-posed questions due to audio or text facial motion information may have multiple corresponding solutions. For instance, different facial poses or motions of same identity can produce basically the same audio. Cross-modality face reenactment is challenged to attract attention, which will also be introduced comprehensively. Text driven methods are developed based on three stages in terms of driving content progressively, which are extra required audio, restricted text-driven and arbitrary text-driven. The audio driven methods can be further divided into two categories depending on whether additional driving information is demanded or not. The additional driving information refers to eye blinking label or head pose videos, which offer auxiliary information in generation procedure. Moreover, comparative experiments are conducted to evaluate the performance between various methods. Image quality and facial motion accuracy are taken into consideration during evaluation. The peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), cumulative probability of blur detection (CPBD), frechet inception distance (FID) or other traditional image generation evaluation metrics are adopted together. To judge the facial motion accuracy, landmark difference, action unit detection analysis, and pose difference are utilized. In most facial-related cases, the landmarks, the presence of action unit or Euler angle are predicted all via corresponding pre-trained models. As for audio driven methods, the lip synchronization extent is also estimated in the aid of the pretrained evaluation model. Apart from the objective evaluations, subjective metrics like user study are applied as well. Furthermore, the commonly-used datasets in face reenactment are illustrated, each of which contains face images or videos of various expressions, view angles, illumination conditions or corresponding talking audios. The videos are usually collected from the interviews, news broadcast or actor recording. To reflect different level of difficulties, the image and video datasets are tested related to indoor and outdoor scenario. Commonly, the indoor scenario refers to white or grey walls while the outdoor scenario denotes actual moving scenes or the news live room. As for conclusion part, the practical applications and potential threats are critically illustrated. Face reenactment can contribute to entertainment industry like movie video dubbing, video production, game character avatar or old photo colorization. It can be utilized in conference compressing, online customer service, virtual uploader or 3D digital person as well. However, it is warning that misused face reenactment behaviors of lawbreakers can be used for calumniate, false information spreading or harmful media content creation in DeepFake, which will definitely damage the social stability and causing panic on social media. Therefore, it is important to consider more ethical issues of face reenactment. Furthermore, the development status of each category and corresponding future directions are displayed. Overall, model optimization and generation-scenario robustness are served as the two main concerns. Optimization issue is focused on data dependence alleviation, feature disentanglement, real time testing or evaluation metric improvement. Robustness improvement of face reenactment denotes generate high-quality reenacted faces under situations like face occlusion, outdoor scenario, large pose faces or complicated illumination. In a word, our critical review covers the universal pipeline of face reenactment model, main challenges, the classification and detailed explanation about each category of methods, the evaluation metrics and commonly used datasets, the current research analysis and prospects. The potential introduction and guidance of face reenactment research is facilitated.

Key words: artificial intelligence (AI); computer vision; deep learning; generative adversarial network (GAN); DeepFake; face reenactment

0 引　言

重演即对事件或事物的再现或重现。在计算机视觉领域,重演是指将一个物体的动作迁移到另一个同类物体上。例如机械臂的移动、动物的肢体动作等。面部重演则更为精细,聚焦于人类面部动作姿态的再现,包括但不限于表情、嘴部动作、眼神朝向和头部姿态等。具体而言,给定源人脸图像和驱动人脸的图像或运动视频,面部重演希望将驱动人脸的面部动作姿态迁移到源人脸上,并以对应的图像或视频的方式呈现出来。如图 1 所示,给定源人脸和驱动人脸,重演图像包含了源人脸的身份以及驱动人脸的面部属性信息。随着技术的发展,音频和文本等非图像级别的驱动信息也开始应用于面部重演研究。

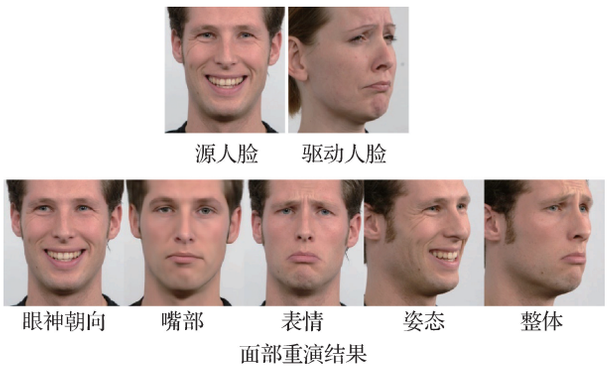


图 1　面部重演的例子和解释

Fig. 1　Examples and illustrations of face reenactment

面部重演有很多新颖的实际应用。例如通过修改嘴部动作使本地化配音后的外国电影嘴型匹配(Prajwal 等,2019)、通过对历史人物的面部重演制作博物馆的宣传教育视频(Lee,2019)、通过对视频会议中人物的面部重演进行会议视频流压缩(Wang 等,2021c)、游戏或虚拟现实中的 3 维人脸建模(Nagano 等,2018)等。尽管面部重演有广泛而实际的应用,这项技术也有其潜在的危害,攻击者可以使用面部重演技术诋毁他人、散布错误信息等。例如通过视频聊天窗口扮演他人以获取其家人信任进行欺诈、生成负面有害内容对他人进行敲诈勒索、篡改他人说话视频以达到不法目的等。

目前,国内外几乎没有重点关注面部重演的综述,相关面部重演的综述内容只是在深度伪造检测综述中以深度伪造的内容出现(Lyu,2020;Tolosana 等,2020a,b;Li 等,2021b;Mirsky 和 Lee,2022)。本文重点关注面部重演生成本身,详细介绍面部重演模型,涵盖了问题定义、方法分类、现状与挑战等内容。因此,本文是上述综述文献很好的补充。

本文的内容框架如图 2 所示。根据驱动信息模态的不同,将面部重演分为图像驱动面部重演和跨模态驱动面部重演,并在图像驱动面部重演中根据驱动特征表达的不同进行细粒度的分类阐述。具体而言,本文从面部重演模型的方法概述入手,首先对面部重演方法进行分类、概述和分析。接着对面部重演中模型训练常用的评价指标及数据集进行列举、说明,最后对本领域进行总结与展望,探究演化趋势、当前挑战、未来发展方向、潜在的危害及应对策略等。

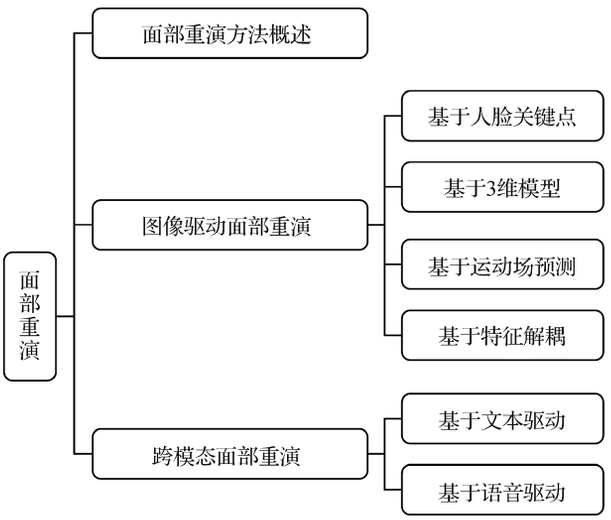


图 2　本文主要内容框架

Fig. 2　The framework of this paper

1 面部重演方法概述

给定源人物的人脸图像以及驱动信息,面部重演目标是生成对应的人脸图像或视频,生成的媒体内容包含源人物的身份信息以及驱动信息对应的嘴部动作、表情、眼神朝向和头部姿态等。此外,背景、光照信息等要与源人物图像中的保持一致。

面部重演模型整体流程如图 3 所示。编码器从源人脸图像提取身份信息和背景信息的表达,表征提取模块从驱动信息中提取驱动信息的表达,两种表达送入生成器得到最终的面部重演图像。

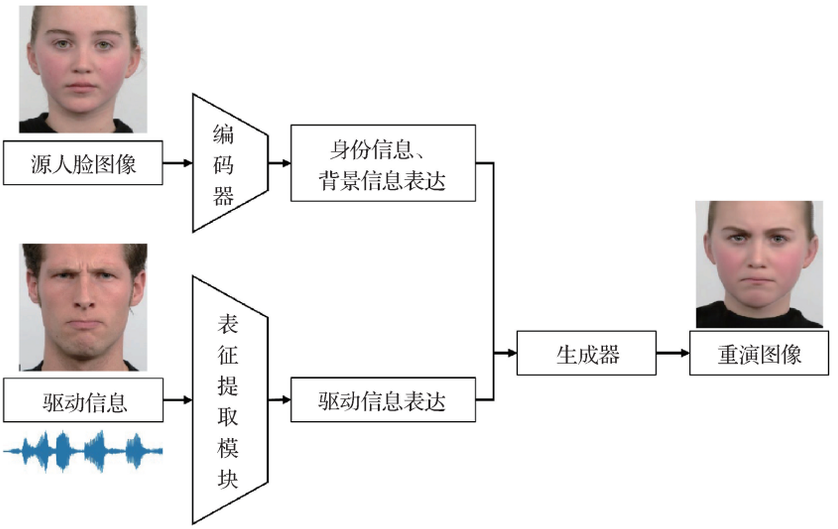


图 3　面部重演流程图

Fig. 3　The flowchart of face reenactment model

1.1　人脸特征表达

在面部重演中,人脸特征表达分为身份信息表达和驱动信息特征表达两种。

身份信息表达通常是在面部重演网络结构中直接采用预训练的人脸识别模型提取特征,或在生成器中将源人脸图像作为输入提供身份信息。身份信息表达方法大同小异,不同的面部重演模型的主要变化在于驱动信息特征的表达方式。

驱动信息特征表达在面部重演中至关重要,因为生成的图像需要包含从驱动信号提取到的面部动作信息。图 4 展示了人脸驱动信息表征示意图,其中面部动作单元示意图修改自面部动作单元分析的综述(Martinez 等,2019)。常用驱动信息表征手段包括向量、人脸关键点、面部边界、面部动作单元和 3 维形变模型(Blanz 和 Vetter,1999)等。向量指将含有相同面部姿态的人脸放入神经网络提取的相同特征。人脸关键点是指事先定义好的位于面部轮廓、眼睛、鼻子和嘴巴等的位置点,常用单通道灰度图表示。面部边界以多幅热力图的方式呈现,通常手动将面部分为几个重要区域,例如上眼睑、鼻梁

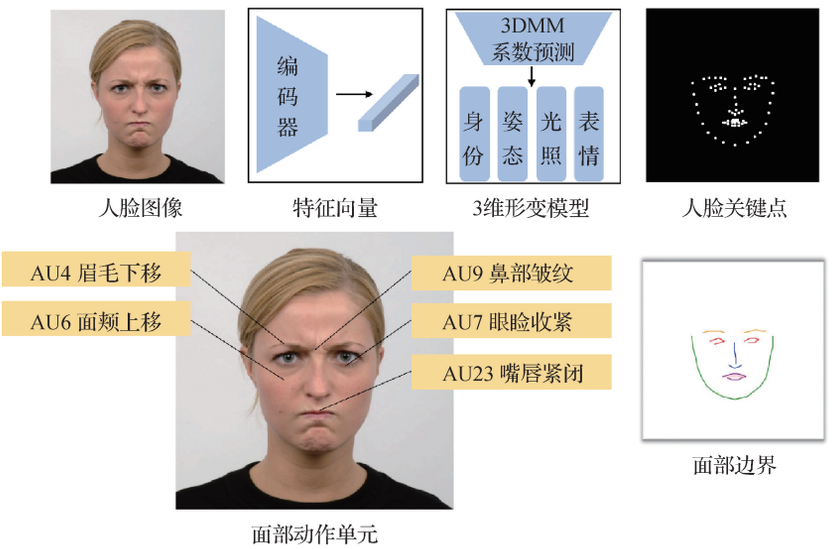


图 4　面部重演驱动信息表达示意图

Fig. 4　Samples of driving information representation in face reenactment model

等。每幅热力图对应一个区域的轮廓信息。3 维形变模型是通过对 2 维人脸图像的单目重建得到的 3 维模型，其中表情、脸部朝向和光照等姿态信息转化为 3 维形变模型的参数。

1.2 面部重演的问题

面部重演发展历程中常见的问题和挑战主要包括身份泄露、时序不一致性、人脸背景的不一致和身份限制。

1）身份泄露。从面部重演的定义可知，面部重演模型最终得到的重演图像的面部姿态信息应当与驱动图像一致，身份信息应当与源图像一致。身份泄露指重演图像中包含了驱动人脸图像的身份信息。该现象的成因在于未能有效分离驱动信号中包含的其他身份信息，对最终的重演生成造成了干扰。常见的解决方法包括对驱动信号进行解耦（Ji 等，2021）或者后处理驱动信息表达特征（Liu 等，2021）。

2）时序不一致性。面部重演模型的输出有图像和视频两种，当考虑到视频级别的生成时，某些重演视频可能会产生伪影、抖动或不规则的头部转动等，这种情况称为时序不一致性。这种现象的成因在于模型训练时是以图像为粒度进行建模，忽略了对图像时序级别的约束和建模。常见的解决方法包括使用循环神经网络进行生成（Suwajanakorn 等，2017），生成器会考虑过往生成的图像帧，判别器也会综合考虑连续帧而并非单独的一帧（Huang 等，2020）。

3）人脸背景的不一致。传统的面部重演模型训练使用的数据集是实验室场景拍摄的，背景单一，生成较为容易。随着 VoxCeleb 和 VoxCeleb2 数据集的提出，真实场景下的面部重演受到广泛关注。当前景的人脸头部转动时，背景信息的生成需要考虑到周围的变化，未考虑此情况的模型会产生人脸与背景交界处的明显痕迹。通常用图像扭曲或图像修复方法处理人脸和背景的交界处（Siarohin 等，2019b），直接预测黑色背景的重演图像随后进行后处理也是解决方法之一（Burkov 等，2020）。

4）身份限制。传统的面部重演模型在模型训练完成后进行推断时，只能接受与训练时相同的人物身份的源图像，这种限制阻碍了面部重演的泛化性，制约了在实际中的应用范围。该问题成因在于模型没有自适应从源图像提取身份信息和背景信息的能力。主流方法开始关注多对多面部重演，即推断时源图像和驱动图像均无身份限制，通常采用鲁棒的人脸身份或姿态的表达（Zeng 等，2020）、小样本学习（Zakharov 等，2019）以及 3 维建模（Yao 等，2020a）等方法解决。

1.3 面部重演的分类

面部重演模型可以从驱动模态、驱动信息表达方式和推断时的身份限制等 3 方面进行分类。

根据面部重演模型中驱动模态的不同，可以将面部重演分为图像驱动的面部重演和跨模态面部重演。传统的面部重演的驱动模态为图像模态，即使用视觉级别的信息驱动面部重演。与之相对应的是跨模态面部重演，即驱动模态的信息为文字或音频。跨模态面部重演由于不同模态的信息容量无法完全对应而退化为病态问题，如何找到不同模态信息之间的对应关系是跨模态面部重演中的重要问题。

根据驱动信息表达方式的不同，可以将面部重演模型分为基于人脸关键点的方法、基于 3 维模型的方法、基于运动场预测的方法和基于特征解耦的方法。基于人脸关键点的方法从驱动模态中提取人脸关键点、人脸边缘图和面部动作单元等信息作为驱动信息表达，常采用基于图像翻译的生成器进行面部重演。基于 3 维模型的方法从驱动信息中提取 3 维人脸形变模型的信息作为驱动特征，或使用神经辐射场方法对说话人场景进行建模，最后采用神经渲染的方式得到重演图像。基于运动场预测的方法从驱动信息和源人脸中寻找运动场差异信息，随后采取基于遮罩和特征扭曲的生成器进行面部重演。基于特征解耦的方法从驱动信息中提取不同语义空间的特征向量作为驱动特征表达。

根据推断时有无身份限制，可以将面部重演模型分为有身份限制的模型和无身份限制的模型。有身份限制的模型仅可以生成特定人物的重演图像或视频；无身份限制的模型则可以根据源图像的指导生成各种身份的重演图像或视频。

2 图像驱动的面部重演模型

图像驱动的面部重演是指在重演过程中人脸驱动姿态的来源为图像或视频帧。该方法可以直观提取图像模态中的人脸姿态信息，指导人脸进行生成。图像驱动的面部重演可以根据人脸驱动信息表达方式的不同进行细粒度的分类，包括基于人脸关键点的方法、基于 3 维模型的方法、基于运动场预测的方

法和基于特征解耦的方法。本文以每种方法的代表工作为基础,递进式地阐述不同类别方法的研究现状与发展,并对不同种类的图像驱动面部重演方法的代表性工作进行实验比较分析。

2.1 基于人脸关键点的面部重演方法

基于人脸关键点的面部重演的基本框架如图5所示。首先基于驱动人脸得到关键点信息,随后使用源人脸的图像作为生成器的输入,在生成器中插入驱动关键点的特征信息,最终生成器的输出即为重演后的人脸图像。由于模型和训练数据的限制,有些方法在推断时仅可以生成特定身份范围的人脸图像,有些方法则没有身份上的限制。本文从重演推断时有无身份限制进行分类,分别阐述相关的面部重演方法。

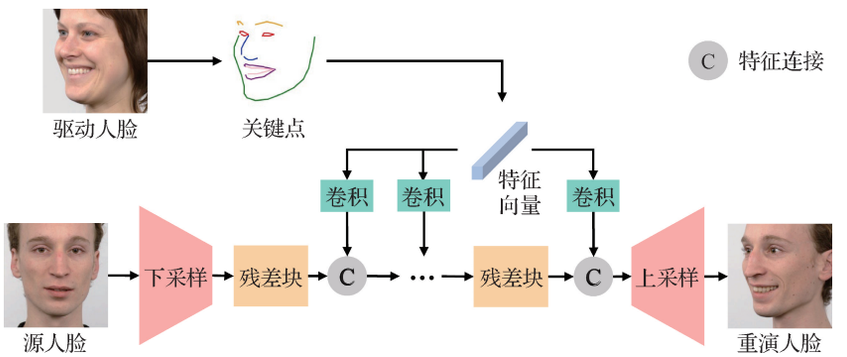


图5 基于人脸关键点的面部重演方法的基本流程图

Fig. 5 Basic flowchart of landmark based face reenactment methods

2.1.1 有身份限制的基于人脸关键点的方法

该类方法的代表性工作由Wu等人(2018)提出。该方法根据驱动人物生成特定人物的重演图像,首先提取驱动人脸的关键点结构信息,并将人脸关键点根据语义进行分块,映射到边界隐空间中,随后变换为特定目标人脸的边界,最后使用基于U-Net的生成器重演特定的目标人脸图像。在损失函数方面,该方法使用传统的生成器判别器损失、重建损失以及特征一致性损失。重建损失的具体表达式为

$$\mathcal{L}_{\mathrm{R}} = \|\boldsymbol{I} - \hat{\boldsymbol{I}}\| \tag{1}$$

式中,$\hat{\boldsymbol{I}}$为模型生成的图像,$\boldsymbol{I}$为与之对应的真实图像,重建损失旨在约束图像像素级别的一致性。特征一致性的损失的具体表达式为

$$\mathcal{L}_{\mathrm{Feat}} = \|V(\boldsymbol{I}) - V(\hat{\boldsymbol{I}})\|_2 \tag{2}$$

式中,V代表VGG-16(Visual Geometry Group 16-layer network)网络的relu2_2和relu3_3层,特征一致性损失旨在约束图像特征级别的一致性。上述3种损失常用于各类重演模型。

Wang等人(2018)专注于视频生成,提出一种基于时间一致性的条件生成对抗网络的面部重演视频生成的方法,在生成当前视频帧时考虑了过往帧,保证了时间一致性。该方法的生成器判别器损失的输入为连续的视频帧,判别器未对整幅图像进行判别,而是对生成图像的不同区域进行判别。此外该方法改进了特征匹配损失,具体表达式为

$$\mathcal{L}_{\mathrm{FM}} = \sum_{i=1}^{T} \frac{1}{N_i}\left[\|D^{(i)}(\boldsymbol{I}) - D^{(i)}(\hat{\boldsymbol{I}})\|_1\right] \tag{3}$$

式中,D为判别器,T为判别器的层数,N_i为每层判别器输出的元素数。Ma等人(2019)改进了生成器的训练范式,借鉴了条件生成对抗网络,根据驱动人脸生成人脸关键点图像,随后训练图像生成模块,将人脸关键点图像转化为特定人物的图像,达到面部重演的效果。Liu等人(2019)对驱动模态生成进行拓展,通过驱动视频预测面部动作单元和人脸关键点,然后根据源人物生成面部轮廓边界,最后使用图像生成模块生成指定源人物的重演人脸图像,提升了面部细节质量。

2.1.2 无身份限制的基于人脸关键点的方法

该类方法的代表性工作为Zhang等人(2020b)提出的FreeNet(multi-identity face reenactment network)。该方法为了解决重演人脸与源人脸身份不匹配问题,采用了基于人脸关键点的两阶段生成方法。首先微调驱动人物人脸关键点以适配源人物的身份信息,随后使用微调后的人脸关键点和源人物的人脸图像进行生成。生成器部分采用经典的"编码器—残差连接层—解码器"架构,源人脸作为生

成器的输入，人脸关键点的特征与残差连接层的特征进行拼接以引入到生成器中，最终得到重演的人脸图像。该方法在训练过程中额外引入三元感知损失，使当前人物的关键点微调过程参与其他人物的面部重演训练，避免了生成网络训练时的模式坍塌。三元感知损失的具体表达式为

$$\mathcal{L}_{\mathrm{TP}} = [\, m + D(\boldsymbol{\kappa}(\hat{\boldsymbol{I}}_{s,e_1}), \boldsymbol{\kappa}(\hat{\boldsymbol{I}}_{s,e_2})) - D'(\boldsymbol{\kappa}(\hat{\boldsymbol{I}}_{s,e_1}), \boldsymbol{\kappa}(\hat{\boldsymbol{I}}_{d,e_1}))\,]_+ \tag{4}$$

式中，m 是控制类内和类间特征的阈值，$\boldsymbol{\kappa}(\cdot)$ 是使用 VGG 网络提取出的特征向量，D' 为两向量间的 L2 距离，下角 + 表示该项的值是非负的，$\hat{\boldsymbol{I}}_{s,e_1}$ 和 $\hat{\boldsymbol{I}}_{s,e_2}$ 分别为源人脸处于表情 1 和表情 2 的图像，$\hat{\boldsymbol{I}}_{d,e_1}$ 为驱动人物处于表情 1 的图像。该损失在特征空间上约束了生成的图像，约束同身份人物图像的特征距离较小而不同人物图像的特征距离较大。此外，该方法同样使用了生成器判别器损失和重建损失。

Sanchez 和 Valstar（2018）则引入了注意力掩码，将人脸关键点图像与源图像送入生成网络中得到重演图像，最后使用注意力掩码将生成的图像与源图像加权得到最终的重演人脸图像。在训练过程中，该模型使用了循环损失，使模型能够从重演人脸图像恢复出源图像，以此增强模型的泛化能力，其表达式为

$$\mathcal{L}_{\mathrm{cycle}} = \| G(G(\boldsymbol{I}, l_d), l_s) - \boldsymbol{I} \| \tag{5}$$

式中，G 为图像生成器，$\boldsymbol{I}$ 为源人脸图像，l_s 和 l_d 分别为源人脸和驱动人脸的关键点。

此外，该模型使用三元循环损失减少图像伪影，即约束两幅不同姿态相同身份的源图像被同一人脸关键点驱动得到的重演人脸图像的一致性。三元循环损失的具体表达式为

$$\mathcal{L}_{\mathrm{triple}} = \| G(\bar{\boldsymbol{I}}, l_d) - G(\boldsymbol{I}, l_d) \| \tag{6}$$

式中，$\bar{\boldsymbol{I}}$ 和 $\boldsymbol{I}$ 为两幅不同姿态但身份相同的源图像，l_d 为驱动人物的关键点。

Pumarola 等人（2018）将面部动作单元作为表情信息的特征表示，使用注意力掩码处理遮挡并减少面部伪影，在模型训练阶段同样采用循环一致性损失进行约束，该方法采用两个不同的判别器判断图像的真伪和预测人脸图像的面部动作单元标签。

为了解决面部重演中的常见问题，相关方法设计独特结构改进重演模型。Nirkin 等人（2019）提出一种可以有效处理脸部遮挡的方法，生成网络接收源人脸视频和驱动视频中的人脸关键点，得到重演后的图像及其分割图，生成过程中使用了 Delaunay 三角分割人脸和质心坐标插值，充分使用源视频中不同的人脸图像提供的信息；为了减少身份不匹配问题，Sun 等人（2021）提出一种重演模型，添加了对重演后的图像的人脸关键点检测，对生成图像进行额外的结构上的约束。Ha 等人（2020）引入人脸关键点转换模块、特征对齐模块和基于注意力机制的特征融合模块，通过接收驱动人脸和多幅源人脸完成面部重演的任务。Tripathy 等人（2021）也采用这种范式，引入动作单元预测模块对转化后图像的人脸关键点进行约束，最后将分割得到的源图像背景和生成图像通过融合模块进行融合，生成最终的重演图像。Liu 等人（2021）发现基于人脸关键点的方法处理大角度人脸时会产生伪影等痕迹，通过采用关键点转换模块、脸部旋转模块和表情增强模块进行多阶段逐步重演，解决了重演人物身份变化和大角度重演效果差的问题。为了解决分辨率低的问题，Fu 等人（2021）提出一种基于人脸边界的高分辨率大角度面部重演方法，首先将源人脸边界的编码、驱动姿态和表情的编码放入解码器生成得到人脸边界图，然后使用人脸边界图和源人脸图像进行面部重演，训练过程中使用不同尺度的重建损失进行约束以保证高分辨率下的重演效果。

此外，为了增加模型的可解释性以及生成时的鲁棒性，在面部重演中分别提取身份信息和姿态信息。Xiang 等人（2020）通过显式地将人脸关键点分离出身份信息和姿态信息，构建特征字典寻找人脸关键点与图像之间的对应关系，通过写操作写入图像的特征信息，同时使用读操作根据关键点生成图像。Wang 等人（2020b）提出一种基于关键点表达解耦身份和动作特征的重演方法，首先自监督预测源图像和驱动图像的关键点以及关键点附近的仿射变换，通过源图像预测外观特征、姿态无关的关键点信息、头部姿态信息和表情形变来完成整体关键点预测，通过驱动图像预测姿态和表情形变，随后该方法通过预测光流场变换来扭曲源特征，通过扭曲后的特征生成重演图像。

2.1.3　小结

当前，人脸关键点检测与定位技术较成熟，基于人脸关键点的面部重演方法可以简单高效地表征人

脸结构信息,但是粗暴地将驱动信息直接映射到人脸关键点等结构信息的做法忽略了驱动信息本身的语义级别特性,例如表情、大角度姿态等。此外,人脸关键点受本身粗粒度的限制,无法表征细节的内容,上述问题均会导致生成人脸结构信息的不准确。

2.2 基于3维模型的面部重演方法

基于3维模型的面部重演方法主要包括人脸的3维形变模型(3D morphable model,3DMM)(Blanz和Vetter,1999)。3DMM可对人脸进行重建或渲染,实现人脸—参数—人脸的转化过程,其中参数为表情、姿态和身份参数,通过直接修改参数进行渲染或借用参数作为人脸特征均在面部重演中有所应用。3DMM模型表达式为

$$\boldsymbol{S}_{\text{model}} = \overline{\boldsymbol{S}} + \sum_{i=1}^{m} \alpha_i \boldsymbol{S}_i \tag{7}$$

$$\boldsymbol{T}_{\text{model}} = \overline{\boldsymbol{T}} + \sum_{i=1}^{m} \beta_i \boldsymbol{T}_i \tag{8}$$

式中,$\overline{\boldsymbol{S}}$和$\overline{\boldsymbol{T}}$分别为平均脸型及平均纹理,用于表示人脸中共性的部分,$\boldsymbol{S}_i$和$\boldsymbol{T}_i$分别为第i张脸的脸型及纹理基,α和β为对应的系数。式(7)用于表达人脸形状,式(8)用于表达脸部纹理。通常情况下,在基确定的情况下,模型会通过预测对应的系数来预测人脸形状。

基于3维模型的面部重演的基本框架如图6所示,该流程修改自Yao等人(2020a)的工作。首先分别对源人脸和驱动人脸进行3DMM的系数回归,得到表情、姿态和身份系数。随后根据重演的定义进行交叉组合,得到新的驱动3DMM系数。然后进行3维人脸重建,得到源人脸的3维模型和新的驱动人脸的3维模型,将上述两种3维模型和2维平面内的源人脸图像一同放入到重演模块中进行生成,得到最终的重演人脸图像。按推断时有无身份限制,基于3维模型的面部重演方法分为有身份限制的基于3维模型的方法和无身份限制的基于3维模型的方法。

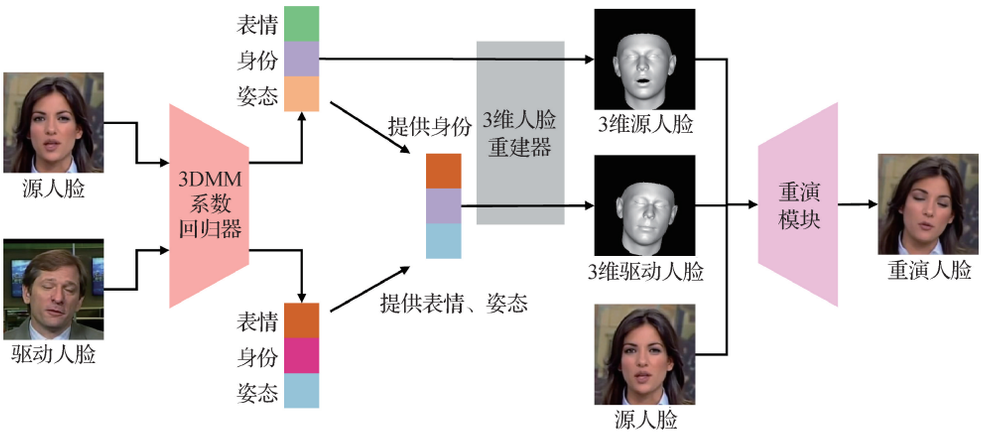


图6 基于3维模型的面部重演方法的基本流程图

Fig. 6 Basic flowchart of 3DMM based face reenactment methods

2.2.1 有身份限制的基于3维模型的方法

Kim等人(2018)提出一种基于时空编码及3维模型的方法,将接收的驱动视频和特定人物的源视频作为输入,对视频中人物进行3维重建,预测光照、身份、姿态、表情和眼部动作参数,随后交叉组合并延展为时序特征输入到渲染模块,得到头部图像、纹理贴图和眼部姿态图像,最后输入到生成器中得到重演视频。

Thies等人(2020)直接从驱动信息中提取特征,预测对应的3DMM的表情参数,根据源人物视频片段预测身份参数和纹理参数,最后使用两个生成网络分别生成新的脸部图像以及将新的脸部图像融合到源图像上,实现面部重演。

2.2.2 无身份限制的基于3维模型的方法

此类方法的代表方法为Thies等人(2015)提出的一种基于RGB深度相机的实时面部重演方法。该方法使用密集跟踪模块接受图像和深度图的输入,输出人脸模型的纹理、反射、光照、姿态和表情参数,通过融合驱动人脸的表情参数和源人脸的其他参数进行人脸3维重建,得到人脸的3DMM模型,随后通过渲染生成2维平面上的人脸,最后辅助以牙齿信息和嘴部图像,融合得到最终的重演图像。

在此基础上，Thies 等人(2016)将视频输入改为仅使用 RGB 相机捕获且添加了嘴部检索模块，从源视频选用与当前驱动人脸最接近的嘴型作为重演后的嘴部区域。

Shen 等人(2018a)专注于重演图像的身份保持的提升，额外引入身份判别器进行身份信息的分类，并引入图像质量判别器。随后，Shen 等人(2018b)继续提升图像生成的多样性，采取多个生成对抗网络生成不同的 3 维人脸模型特征分布，随后在分布上进行均匀采样和插值，最后结合源人物的身份特征放入生成器中进行重演图像的生成。

Nagano 等人(2018)摒弃视频流的源人脸输入，对单幅源图像进行重演。首先从图像中获取头部网格、表情网格、UV 贴图、深度图和纹理信息等，放入 U-Net 网络中生成不同表情的图像，再从这些表情中提取面部动作单元，最后根据驱动表情组合上述面部动作单元完成重演。Thies 等人(2019)提出一种基于噪声纹理图像渲染的重演范式，通过提取驱动人脸的表情信息以及源人脸的身份信息合成 UV 贴图，然后在神经网络学习到的纹理图中进行采样，得到重演图像的纹理信息后渲染出最终的人脸。Yao 等人(2020a)则充分利用了 3DMM 中的节点信息，使用图卷积网络对源人脸和驱动人脸的 3 维模型进行建模，提取光流信息进行重演。

2.2.3　小结

基于 3 维模型的方法可以有效分离出驱动人脸的身份、姿态和表情等特征，提高了面部重演阶段对于人脸的可控性。但是，基于 3 维模型的源人脸交互的驱动特征提取方法会增大模型训练和推断的负担。此外，3 维人脸形变模型并不包含背景信息的建模，所以在生成的人脸中，背景以及前景和背景的衔接处会有明显伪影痕迹，造成了生成质量的下降。

2.3　基于运动场预测的面部重演方法

基于运动场预测的面部重演模型在提取驱动信息特征表达时可以更加充分地利用源人脸像素级别的图像信息。如图 7 所示，通常情况下模型会接受若干幅源人脸图像并预测中立嵌入人脸，中立嵌入人脸保留了源人脸的纹理信息。随后运动场预测模块根据驱动人脸图像预测出运动场，并通过图像扭曲模块作用到驱动人脸上得到最终的重演人脸。基于运动场预测的面部重演方法均为无身份限制的方法，有身份限制的方法在训练时接收到的人脸数据是一个或若干个特定人物的，但是由于运动场预测方法本身检测了源人脸和驱动人脸的结构信息，所以并未对源人脸的身份信息做出限制。

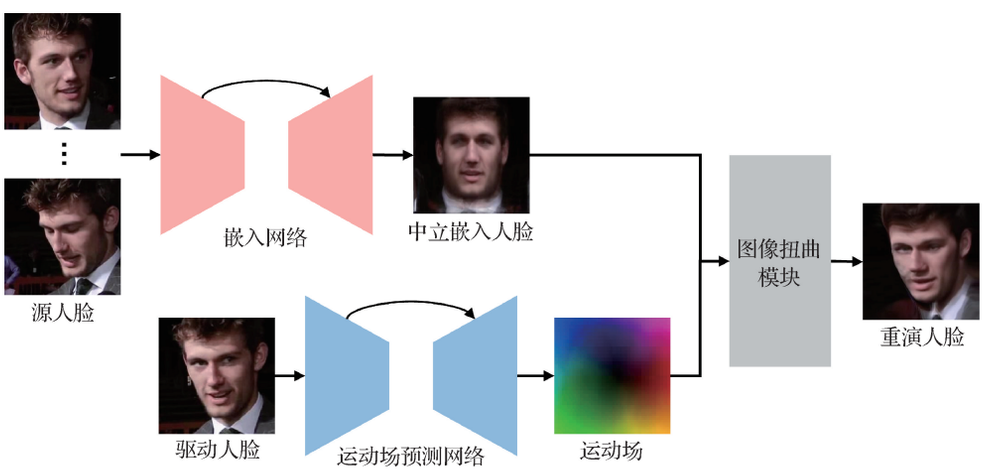


图 7　基于运动场预测的面部重演方法的基本流程图

Fig. 7　Basic flowchart of motion field prediction based face reenactment methods

2.3.1　方法概述

此类方法的代表方法由 Wiles 等人(2018)提出，该方法可以进行基于单幅源图像的面部重演。首先将一幅或多幅源图像映射到中立嵌入人脸图像，该图像充分利用源人脸的结构信息和纹理信息，随后使用卷积神经网络基于驱动信息预测稠密的像素级别的扭曲场，对中立嵌入人脸进行图像扭曲操作即可得到最终的重演人脸。在损失函数方面，该方法使用了生成器判别器损失、逐像素重建损失以及身份损失，其中身份损失对源图像和重演图像进行人

脸识别身份特征的提取,以最小化特征之间的距离。

为了提升运动场预测的准确性,研究者使源图像与驱动图像进行交互,预测运动场并设计基于掩码的生成器完成面部重演。Siarohin 等人(2019a)通过自监督方法预测人脸的结构运动关键点,在运动场预测时引入源人脸的纹理信息。随后,Siarohin 等人(2019b)改进上述方法,提出基于一阶泰勒展开的运动场预测重演方法,使用关键点附近的仿射变换对姿态信息进行建模,生成器部分负责预测遮挡掩码,在面部重演的部分引入图像修复流程,提升人脸前后景衔接处的质量。

不同于上述深度学习领域的卷积预测操作,一些方法引入传统的人脸对齐操作提取运动场信息。Averbuch-Elor 等人(2017)对齐视频中的每一个驱动帧和源人脸图像的脸部结构信息,使用 Delaunay 三角进行区域分割,并采用 2 维图像形变的方法生成粗粒度图像,随后生成嘴部区域的图像以及添加细粒度的细节信息完成重演。Geng 等人(2018)的方法通过人脸关键点对齐后的图像扭曲生成粗粒度图像,通过不同的生成器分别强化图像细节和生成嘴部细节。

与代表方法直接使用解码器进行图像生成类似,研究者同样采取渐进式生成架构或双流生成架构提升图像的生成质量,即在生成器部分进行改进。Zhang 等人(2019)提出一种基于分割图的多尺度重演方法,从源图像提取外观特征向量,从驱动图像提取分割图,生成器接收源图像特征作为输入,逐层将分割图嵌入到生成器中得到粗粒度人脸,最后与通过图像扭曲得到的人脸通过掩码加权融合得到最终的细粒度重演结果。Gu 等人(2020)提出一种端到端的基于纹理和图像扭曲的双流面部重演网络,通过扭曲支路预测运动场和选择掩码,加权得到扭曲人脸,另一支路则负责生成从源人脸中无法通过扭曲直接得到的人脸部分,将上述两个支路通过掩码加权组合即得到最终的重演人脸。Zakharov 等人(2020)提出了双流神经合成的方法,使用纹理生成网络接收源图像以及对应人脸关键点的特征集合,生成高频纹理信息,随后推断阶段接收姿态信息生成低频图像和运动场,运动场负责扭曲纹理信息并添加到低频图像上,最终生成重演后的图像。为提升面部细节的生成质量,Yao 等人(2021)提出局部生成整体优化的方法,采取光流预测和图像扭曲,通过眼、鼻和嘴等部位的局部生成而后融合的方式进行面部重演。

2.3.2 小结

基于运动场预测的方法高效利用了源人脸提供的身份信息和背景信息,保证了一定的面部纹理细节的逼真程度以及背景图像的准确性。然而,基于运动场预测的方法在生成器部分是对源图像进行采样,并通过预测遮罩掩码进行图像生成,所以该方法生成人脸的姿态严重依赖于源人脸的初始姿态,无法生成大角度人脸,并且该种方法忽略了源人脸显式结构信息的提取,会造成生成图像的身份不匹配现象,即生成人脸的身份信息与源人脸的身份信息相比有偏差。

2.4 基于特征解耦的面部重演方法

基于特征解耦的面部重演方法通常不引入额外的人脸关键点或面部动作单元等结构信息,而是直接对人脸图像的运动和身份信息进行显式地强解耦。基于特征解耦的面部重演方法的基本流程如图 8 所示,该流程图修改自 Burkov 等人(2020)的工作。在模型训练时,身份编码器接收若干幅源人脸图像的输入,提取若干个身份编码,取平均后作为身份特征。姿态编码器则接收增广后的驱动人脸图像,目的是有效提取驱动人脸的结构信息,得到驱动特征。身份特征和驱动特征融合后,通过全连接层进行拼接,并输入到生成器中完成重演人脸的生成。在生成器中,常用规范化方法和卷积参数预测方法等融合身份和驱动特征。

基于特征解耦的面部重演方法同样为无身份限制的,在模型训练和推断时均提取了源人脸和驱动人脸的信息,未对源人脸数据的身份信息做出额外的限制。

2.4.1 方法概述

该类方法的代表方法由 Zakharov 等人(2019)提出。该方法直接从若干幅源图像中提取身份特征向量,从驱动图像中提取姿态特征向量,将姿态特征嵌入到以身份特征作为输入的生成器中得到重演图像。在训练阶段使用传统的生成器判别器损失、重建损失和特征匹配损失,可以有效分别从驱动图像和源图像中提取模型所需的特征,降低了模型的设计难度,提升了对于大规模人脸数据的需求。在此基础上,Burkov 等人(2020)改进了驱动信息的提取,提出了基于隐空间姿态描述子提取的方法,对驱动图像进行数据增广,使其保留人脸的轮

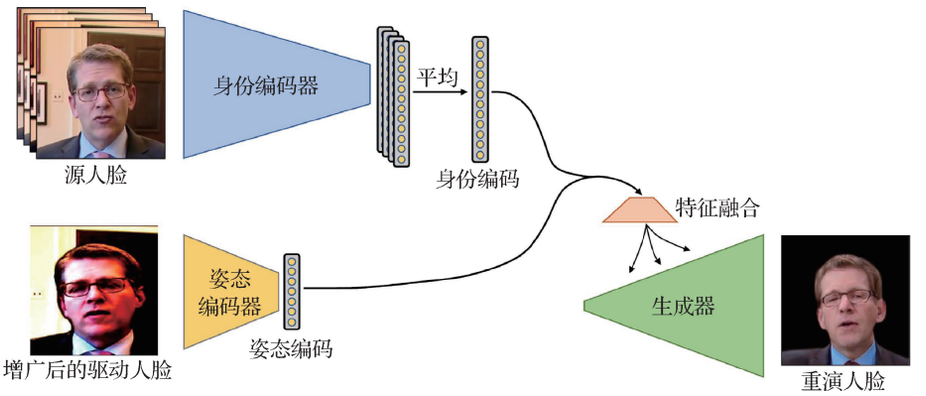


图 8　基于特征解耦的面部重演方法的基本流程图

Fig. 8　Basic flowchart of feature decoupling based face reenactment methods

廓信息，采用姿态编码器提取较低维度的姿态特征。

一些方法尝试在驱动特征提取上进行改进以准确描绘驱动人脸的信息。Wang 等人（2020b）提出一种基于时空生成对抗网络的面部重演视频生成方法，将表情的标签作为姿态信息，辅助以随机变量，通过 3 维反卷积生成重演图像。Zeng 等人（2020）提出基于自监督分离身份和姿态信息的面部重演方法，使用编码器解码器完成图像重建任务来提取中间层向量作为姿态编码，在重演时接收多幅源图像，聚合生成可以编码身份信息的中立人脸作为后续生成网络的输入，随后使用自适应示例正则化将姿态编码嵌入生成网络预测位移场和注意力掩码，最终生成重演的图像。

此外，在提取到初步的驱动特征信息表达后，部分方法采用后处理操作剥离驱动特征中的无关属性，以提升重演图像的质量。Huang 等人（2020）对驱动图像提取姿态编码后，采用身份分类器对其进行约束，使姿态编码不含任何身份信息，并采用成对图像判别器和序列图像判别器分别约束图像的真实性和图像间的时间一致性。

2.4.2　小结

基于特征解耦的面部重演方法规范化强、训练成本低，并以极小的代价衔接当前主流的生成器模型进行面部重演。例如 Pix2Pix（Isola 等，2017）、StyleGAN2（Karras 等，2020）等。但是，这些方法忽略了图像模态人脸本身带有的结构信息，造成了不同语义对应特征的杂糅，直接导致了生成结果结构信息和语义信息的不准确。

2.5　实验比较

不同种类的图像驱动面部重演方法的实验比较如表 1 所示，其中部分实验结果引用自 Yao 等人（2020a）的工作。实验使用 VoxCeleb1 数据集，图像质量评价指标为结构相似性（structural similarity，SSIM）和峰值信噪比（peak signal-to-noise ratio，PSNR）。可以看出，1）基于特征解耦的方法忽略了

表 1　图像驱动面部重演方法比较

Table 1　Comparison results on image-driven face reenactment methods

方法	CSIM ↑	SSIM ↑	PSNR ↑	PRMSE ↓	AUCON ↑
Wiles 等人（2018）	0.689	0.719	22.537	3.26	0.813
Zakharov 等人（2019）	0.229	0.635	20.818	3.76	0.791
Ha 等人（2020）	0.755	**0.744**	23.244	**3.13**	0.825
Siarohin 等人（2019b）	0.813	0.723	30.182	3.79	0.886
Yao 等人（2021）	**0.823**	0.73	30.285	3.26	0.831
Yao 等人（2020a）	0.822	0.739	**30.394**	3.20	**0.887**

注：加粗字体表示各列最优结果，↑表示值越大越好，↓表示值越小越好。

人脸的结构信息而直接提取特征向量,人脸身份特征的余弦相似度(cosine similarity,CSIM)较差;2)基于运动场预测的方法使用了图像扭曲操作,难以处理人脸姿态变化场景,人脸姿态均方根误差(root mean square error of the head pose,PRMSE)处于较低水准;3)基于人脸关键点的方法保持了高水准的人脸结构信息,在结构相似性和姿态一致性方面超越了其他种类的方法,但由于关键点粒度较粗,面部动作单元的识别比例(ratio of identical facial action unit values,AUCON)有待提高;4)基于3维模型的方法总体来说取得了较好效果,但是由于建模难度大、推断成本高,在实际应用中存在一定的局限性。

3 跨模态面部重演方法

跨模态面部重演模型的研究方兴未艾,常见的形式为说话人生成,广泛应用于数字人生成的诸多应用领域。跨模态面部重演模型指面部重演模型中的驱动信息由其他模态的信息提供,而身份信息的来源则保持不变。跨模态的驱动信息来源主要包括文字模态和音频模态。

3.1 基于文字驱动的面部重演

基于文字驱动的面部重演使用文本模态驱动源人脸的图像或视频的生成,通常的表现形式为驱使源人物讲述对应的文本内容,本质是从文字序列到源人物视频序列的映射。然而对同一段文本,不同人有不同的讲述方式。即使是同一个人讲述同一段文本,也会因为语气、情感等因素而对应不同的视频序列。因此,文字序列实际对应的并不是某一个确定的视频序列,而是由所有合理的视频序列构成的视频序列分布。

由于视频序列的超高维度,直接使模型学习到视频序列分布是非常困难的。考虑到这个问题,目前基于文字驱动的面部重演方法往往在文本序列的基础上或加入一定的先验知识或加入使用上的限制,用以简化模型的学习过程。

3.1.1 方法概述

Yu等人(2019)在文字的基础上,额外加入了音频作为面部重演中的驱动信息。该方法首先将文字及音频序列输入到长短期记忆网络(long short time memory, LSTM)用以预测嘴部关键点,然后将源人脸转化成素描图。最后,嘴部关键点和素描图一同输入基于条件对抗网络的生成器中生成最终的源人物视频序列。该方法将原本文字到视频的映射简化为文字到嘴部关键点的映射,同时增加了语音的约束,得到了不错的重演效果。Fried等人(2019)提出了通过文字编辑视频的重演方法,在原有视频的基础上,用户对说话内容进行添加、替换和删除操作,该方法能生成逼真的对应视频。由于用户只能对说话内容进行有限的修改,因此极大地简化了模型的输入。同时,利用已有的视频,模型需要做的是在已有视频的基础上进行检索,挑选最适配于用户修改的片段,大幅简化了模型的学习过程。该方法首先利用视频提供的语音素材对齐驱动文本信息,包括添加、删除和重排单词,随后对视频中的每一帧进行3维人脸重建,提取每一帧的身份和姿态参数,并根据语料的排序信息重排表情参数,最后使用渲染模块结合视频背景信息完成重演。Yao等人(2020b)进一步扩展了上述方法,引入参考人物音素及视频,通过建立音素索引、减少搜索空间的方式提升音频音素查找速度;通过引入参考任务表情参数到目标人物表情参数的转化模块,在搜索因素时仅搜索参考人物即可,摆脱了原有重演模型的身份限制。这类通过文字编辑视频的重演方法,通过简化模型的学习目标,得到了不错的重演结果。但因为输入的文字只能进行有限的修改且需要使用大量源人脸数据,在使用上有很大的限制。Li等人(2021a)的方法降低了对源人脸数据的依赖,通过输入与时间对齐好的文本数据作为驱动信息,分别预测头部姿态、面部表情和嘴部形状等信息,随后重建人脸3维模型并使用视频生成器生成面部重演视频。

3.1.2 小结

从需要额外的音频信息作为驱动的重演任务到限制输入文字序列内容的重演,再到任意文本输入的重演,基于文字驱动的面部重演任务越来越复杂。然而,目前基于文字驱动的面部重演仍处于兴起阶段,针对任意文本输入的重演,研究的主要关注点是如何处理文本驱动模态的输入、如何从文本信息中挖掘潜在的时序信息指导面部重演生成,以及如何建立文本模态到图像模态的联系和映射。

3.2 基于音频驱动的面部重演

基于音频驱动的面部重演与基于文本驱动的面部重演一样,本质上是不适定问题,因为语音模态与

图像模态不同,无法提供眨眼、脸部姿态和表情等信息,所以根据有无额外辅助图像模态驱动信息将基于音频驱动的面部重演分为两类。无额外驱动信息的方法中,相关隐式属性常需要从音频中推断得到。

3.2.1 无额外驱动信息的音频驱动面部重演

该类方法的代表工作由 Chung 等人(2017)首次提出。该方法使用一个音频编码器提取音频的特征作为运动信息,使用基于卷积神经网络的图像编码器提取源图像的特征作为身份信息,将两种信息融合后使用图像解码器生成重演后的人脸图像,最后生成的图像经过一个去模糊模块以提升图像质量。损失函数方面,该方法仅使用了重建损失用于约束模型的训练。

Chen 等人(2018)专注于嘴部的重演,整体流程与代表工作相同,改进之处在于引入了音频衍生编码器和光流编码器,分别提取音视频的长序列信息,通过提升特征相似度进而提高重演视频的音视频的一致性。Song 等人(2019)引入了判别器提升图像质量,该方法通过不同的编码器提取源图像和音频片段的特征,将特征组合后作为循环神经网络(recurrent neural network,RNN)每个节点的输入,将每个节点的输出通过图像解码器生成重演图像,同时采用了多个判别器分别负责图像的真实性、视频级别的连贯性以及口型对齐与否。Zhou 等人(2019)在特征空间进行信息解耦,引入判别器和分类器使用对抗训练及对比损失解耦出身份特征空间及说话内容特征空间,并使用对应的编码器从人脸图像和音频中提取对应的特征,然后将两种特征拼接后送入生成器进行生成。Prajwal 等人(2020)使用海量数据预训练了音视频一致判别器,将音频、视频和去掉嘴部的视频输入到模型中进行生成,使用常规图像质量判别器和预训练的音视频一致判别器辅助生成。

上述模型直接使用编码器提取特征向量,生成重演图像或在原有的视频中修改嘴型,没有引入人脸结构信息,造成嘴部动作生成的不准确。因此研究者提出引入人脸关键点信息作为驱动信息表达解决音频驱动面部重演问题。代表工作为 Chen 等人(2019)提出的 AT-VGNet(audio transformation and visual generation network),该方法可以基于图像生成说话人视频。首先使用音频到人脸关键点的 LSTM 网络逐帧生成人脸关键点,随后使用关键点到图像的时序生成网络,引入注意力掩码,根据源图像和关键点生成重演后的人脸图像;此外 AT-VGNet 还引入了基于回归的判别器,在判别的过程中添加了图像重建任务,分别判别图像级别和视频级别的真实性。损失函数方面,使用了传统的生成器判别器损失以及逐像素的重建损失。

Suwajanakorn 等人(2017)和 Jalalifar 等人(2018)的工作则专注于特定人物的音频驱动面部重演,基于奥巴马的说话视频,通过驱动音频得到嘴部的稀疏关键点形状,通过生成网络得到对应的奥巴马嘴部纹理,并添加高频信息来精细化嘴部以及牙齿,最后将生成的嘴部融合到原有视频中。Song 等人(2022)则专注于音频中身份信息的剥离,该方法预处理音频特征,去除身份信息并提取 3DMM 的表情特征,与源视频中的人脸纹理特征和姿态特征进行融合并 3 维重建,随后送入渲染网络中生成嘴部信息,再融合到源视频中生成图像。Zhou 等人(2020)提出的 MakeItTalk 模型同样专注于音频信息的预处理,提升了模型对于身份信息一致性的保持能力。该方法对音频提取内容编码和身份编码,分别预测对应当前音频片段的内容相关和身份相关的人脸关键点的位移,并将上述两种位移施加到源图像对应的关键点上,通过图像生成模块达到音视频配准的效果。Wang 等人(2021a)提出的 Audio-2Head 则引入了音频信息到姿态信息的映射模块,使用 LSTM 动态预测音频对应的头部姿态,使用基于运动场预测的生成范式进行音频驱动面部重演。Wang 等人(2021b)在学习阶段显示使用音频音素信息以及源人物姿态信息进行输入,通过学习特定人物的音频与嘴型的匹配信息并泛化至其他的源图像,最后通过预测动作场完成图像的生成。此外,Zhang 等人(2021)提出了高分辨率音频驱动面部重演方法,该方法通过将 3 维模型与运动场预测方法结合,通过自行收集的网络高分辨率说话人视频进行训练。

3.2.2 有额外驱动信息的音频驱动面部重演

有额外驱动信息的音频驱动面部重演通常除了驱动音频外,还会引入眨眼标签、姿态图像等辅助信息,提升重演结果的多样性和真实性。代表工作为 Zhou 等人(2021)提出的 PC-AVS(pose-controllable audio visual system)模型,该模型为基于隐式模块化音视频特征表达的多阶段算法,引入人脸姿态视频作为辅助驱动信息,目的是使生成的重演视频包含多样的人物姿态。首先使用身份编码器提取身份信

息,随后使用数据增广的手段将音频对应的视频映射到无身份信息的特征空间,然后将该特征映射到说话内容空间和姿态特征空间,同时将音频同样映射到相同的说话内容空间,最后将姿态信息、身份信息和说话内容放入生成器中进行面部重演。

Zhang 等人(2020a)额外引入了眨眼信息的标签,将人脸图像特征、音频信息、眨眼信息和头部姿态信息进行耦合然后编码,生成特定人物的人脸关键点,随后输入到特定人物生成网络中得到特定人脸的重演图像。在此基础上,为了解决生成时的身份限制,Zhang 等人(2020c)提出基于自适应卷积模块的实时面部重演模型,将音频信息、眨眼信息和头部姿态信息通过该模块融合到编码器解码器结构中进行重演。

Wang 等人(2020a)和 Ji 等人(2021)额外引入了表情信息。前者引入额外的表情标签以及对应的表情幅度标签,通过条件生成网络生成带有表情的说话人图像。后者基于特定的音频数据显式地分离音频中的表情信息,该方法通过交叉重建引入重建损失、分类损失和特征损失,将音频特征解耦到内容空间和表情空间,然后预测人脸关键点并进行单目3维人脸重建,得到姿态、表情和纹理参数,组合后投影得到人脸边缘图,放入生成器得到带有表情信息的重演图像。

3.2.3 小结

基于音频驱动的面部重演成为面部重演在当前互联网环境下的一个崭新表现形式。由于本身音频驱动信息包含了显式的时间序列信号,所以生成音唇一致的说话人视频成为基于音频驱动面部重演最直接的呈现手段。为了提升音频驱动面部重演的逼真程度,当前的方法主要集中在对于人脸显式属性和隐式属性的建模。显式属性是指与音频直接对应的嘴唇动作的属性,隐式属性则是指姿态、眨眼和表情等与音频不直接相关的属性。

一些音频驱动面部重演方法在 VoxCeleb2 数据集上的实验结果比较如表 2 所示,部分实验结果引用自 Wang 等人(2021b)的工作。其中评价指标包含累计模糊检测概率(cumulative probability of blur detection, CPBD)、人脸关键点距离(landmark distance, LMD)、音视频偏移程度(audio visual offset, AVOff)和音视频一致置信度(audio visual lip-sync confidence, AVConf)等指标。

表 2 音频驱动面部重演方法比较

Table 2 Comparison results on audio-driven face reenactment methods

方法	FID ↓	CPBD ↑	LMD ↓	AVOff→0	AVConf ↑
Chen 等人(2019)	67.6	0.061	0.382	0.20	4.30
Prajwal 等人(2020)	20.3	0.541	0.273	−2.92	6.65
Zhou 等人(2020)	23.0	0.55	0.482	−2.83	4.38
Wang 等人(2021a)	22.4	0.532	0.314	0.50	2.47
Wang 等人(2021b)	**19.4**	**0.564**	0.252	**−0.08**	**6.98**
Zhang 等人(2021)	21.2	0.559	0.283	0.491	4.54
Zhou 等人(2021)	27.6	0.514	**0.251**	−3.00	6.83

注:加粗字体表示各列最优结果,↑表示值越大越好,↓表示值越小越好,→0 表示值越趋于 0 越好。

从表 2 可以得知,Chen 等人(2019)提出的方法音视频一致性较强,但是由于只生成了人脸内部区域,所以图像质量较差。与此相比,Wang 等人(2021b)提出的音频驱动面部重演方法在图像质量、音视频一致性上都取得了良好效果。

3.3 小结

面部重演方法汇总如表 3 所示。与传统的面部重演相比,跨模态面部重演引入了不同模态的驱动信息,例如文本、音频等,并包括不同的辅助驱动信息,例如眨眼与否、头部姿态以及人脸表情等;在模型结构上,同样依赖于编码器的特征提取或人脸结构信息的引入。

跨模态面部重演应用前景广阔,将数字媒介不同的模态联系到了一起,其发展瓶颈在于包含多种辅助信息的人脸视频数据集较少,没有可以直接反映重演图像与驱动模态信息一致性的评价指标,问

表 3　面部重演方法汇总

Table 3　Summary of face reenactment models

分类	方法	重演				特征表达						模型输入		模型输出		
		嘴部	表情	姿态	眼神	动作单元	3 维模型	运动场	特征向量	关键点	边界	驱动	源	图像	视频	分辨率/像素
基于人脸关键点	ReenactGAN(Wu 等,2018)	√	√	√	–	–	–	–	–	–	√	图像	图像	√	–	256×256
	Vid2Vid(Wang 等,2018)	√	√	√	–	–	–	–	–	–	√	视频	–	–	√	512×512
	AFR-GAN(Ma 等,2019)	√	√	√	–	–	–	–	–	√	–	图像	–	√	–	512×512
	VS-GAN (Liu 等,2019)	√	√	√	–	√	–	–	–	–	√	视频	–	√	–	256×256
	GANnotation(Sanchez 和 Valstar,2018)	√	√	–	–	–	–	–	–	√	–	图像/关键点	图像	√	–	128×128
	GANimation(Pumarola 等,2018)	√	√	–	–	√	–	–	–	–	–	图像	图像	√	–	128×128
	FSGAN(Nirkin 等,2019)	√	√	√	–	–	–	–	–	√	–	图像/关键点	图像	√	–	256×256
	OIPR-Net(Xiang 等,2020)	√	√	√	√	–	–	–	–	√	–	图像/关键点	图像	√	–	256×256
	FreeNet(Zhang 等,2020b)	√	√	–	–	–	–	–	–	√	–	图像/关键点	图像/关键点	√	–	512×512
	LandmarkGAN(Sun 等,2021)	√	√	√	–	–	–	–	–	√	–	图像/关键点	图像/关键点	√	–	256×256
	MarioNETte(Ha 等,2020)	√	√	√	–	–	–	–	–	√	–	图像	图像	√	–	256×256
	FV-NTH(Wang 等,2021c)	√	√	√	–	–	√	–	–	–	–	图像	图像	–	√	512×512
	FACEGAN(Tripathy 等,2021)	√	√	–	–	√	–	–	–	√	–	图像/动作单元	图像/关键点	√	–	256×256
	HFM-Net(Fu 等,2021)	√	√	√	–	–	–	–	–	–	√	图像/标签	图像	√	–	1 024×1 024 *
	LI-Net(Liu 等,2021)	√	√	√	–	–	–	–	–	√	–	图像/关键点	图像	√	–	256×256
基于 3 维模型	DVP(Kim 等,2018)	√	√	√	√	–	√	–	–	–	–	视频	视频	√	–	1 024×1 024 *
	NVP(Thies 等,2020)	√	–	–	–	–	√	–	–	–	–	视频	视频	–	√	512×512
	RET(Thies 等,2015)	√	√	√	–	–	√	–	–	–	–	视频	视频	–	√	1 280×720 *
	Face2Face(Thies 等,2016)	√	√	√	–	–	√	–	–	–	–	视频	视频	–	√	1 280×720 *
	FaceID-GAN(Shen 等,2018a)	√	√	√	–	–	√	–	–	–	–	图像	图像	–	–	128×128
	FaceFeat-GAN(Shen 等,2018b)	√	√	√	–	–	√	–	–	–	–	隐空间变量	图像	√	–	128×128
	PaGAN(Nagano 等,2018)	√	√	√	√	–	√	–	–	–	–	图像	图像	√	–	512×512
	DNR(Thies 等,2019)	√	√	√	–	–	–	–	–	–	–	图像	图像	√	–	512×512
	MOFR-Net(Yao 等,2020a)	√	√	√	–	–	√	–	–	–	–	图像	图像	√	–	256×256
基于运动场预测	BPL(Averbuch 等,2017)	√	√	–	–	–	–	√	–	√	–	图像	图像	√	–	600×800
	wgGAN(Geng 等,2018)	√	√	–	–	–	–	√	–	√	–	图像	图像	√	–	128×128
	X2Face(Wiles 等,2018)	√	√	√	–	–	–	√	–	–	–	图像	图像	√	–	256×256
	OFR-Net(Zhang 等,2019)	√	√	√	√	–	–	√	–	–	√	图像/关键点	图像	√	–	256×256
	Monkey-Net(Siarohin 等,2019a)	√	√	–	–	–	–	√	–	√	–	图像/关键点	图像	√	–	64×64
	FOMM(Siarohin 等,2019b)	√	√	√	–	–	–	√	–	√	–	图像	图像	√	–	256×256
	FLNet(Gu 等,2020)	√	√	√	–	–	–	√	–	√	–	图像/关键点	图像	√	–	224×224
	Fast-Bilayer(Zakharov 等,2020)	√	√	√	–	–	–	√	–	√	–	图像/关键点	图像/关键点	√	–	256×256
	OFR-AAN(Yao 等,2021)	√	√	√	–	–	–	√	–	√	–	图像	图像	√	–	256×256
基于特征解耦	NTH(Zakharov 等,2019)	√	√	√	–	–	–	–	√	√	–	图像/关键点	图像	√	–	256×256
	ImaGINator(Wang 等,2020b)	√	√	–	–	–	–	–	–	–	–	表情标签	图像	–	√	64×64
	DAE-GAN(Zeng 等,2020)	√	√	√	–	–	–	–	–	–	–	图像	图像	√	–	256×256
	CrossID-GAN(Huang 等,2020)	√	√	√	–	–	–	–	√	√	–	视频	图像	√	–	128×128
	NHR-Net(Burkov 等,2020)	√	√	√	–	–	–	–	√	–	–	图像	图像	√	–	256×256
文本驱动	MATV-Net(Yu 等,2019)	√	√	–	–	–	–	–	–	√	√	音频/文本	图像	√	–	256×256
	TET(Fried 等,2019)	√	–	–	–	–	√	–	–	–	–	文本	视频	–	√	1 280×720 *
	I-TET(Yao 等,2020b)	√	–	–	–	–	√	–	–	–	–	文本	视频	–	√	1 280×720 *
	WAS(Li 等,2021a)	√	√	√	–	–	√	–	–	–	–	文字	图像	√	–	512×512

续表 3　面部重演方法汇总

Table 3　Summary of face reenactment models

分类	方法	重演				特征表达						模型输入		模型输出		
		嘴部	表情	姿态	眼神	动作单元	3维模型	运动场	特征向量	关键点	边界	驱动	源	图像	视频	分辨率/像素
音频驱动	YST(Chung 等,2017)	√	–	–	–	–	–	–	√	–	–	音频	图像	√	–	256×256
	LMG(Chen 等,2018)	√	–	–	–	–	–	√	√	–	–	音频	图像	√	–	256×256
	S-Obama(Suwajana 等,2017)	√	–	–	–	–	–	–	–	√	–	音频	视频	–	√	2 048×1 024*
	SFR-GAN(Jalalifar 等,2018)	√	–	–	–	–	–	–	–	√	–	音频	–	–	√	128×128
	TFG-Net(Song 等,2019)	√	√	–	–	–	–	–	–	–	–	音频	图像	–	√	128×128
	DAVS(Zhou 等,2019)	√	–	–	–	–	–	–	√	–	–	音频	图像	√	–	256×256
	Wav2Lip(Prajwal 等,2020)	√	–	–	–	–	–	–	√	–	–	音频	视频	–	√	256×256
	ATVG-Net(Chen 等,2019)	√	–	–	–	–	–	–	–	√	–	音频	视频	–	√	128×128
	EBT(Song 等,2022)	√	√	–	–	–	√	–	–	–	–	音频	视频	–	√	1 920×1 080*
	MakeltTalk(Zhou 等,2020)	√	√	√	–	–	–	–	–	√	–	音频	图像	√	–	256×256
	APB2Face(Zhang 等,2020a)	√	√	√	–	–	–	–	–	√	–	音频	–	√	–	256×256
	APB2FaceV2(Zhang 等,2020c)	√	√	√	–	–	–	–	–	√	–	音频	图像	√	–	256×256
	MEAD(Wang 等,2020a)	√	√	–	–	–	–	–	√	√	–	音频/标签	图像	√	–	256×256
	EVP(Ji 等,2021)	√	√	–	–	–	√	–	–	–	√	音频	图像	√	–	256×256
	PCAVS(Zhou 等,2021)	√	–	√	–	–	–	–	√	–	–	音频/视频	图像	√	–	256×256
	Audio2Head(Wang 等,2021a)	√	–	√	–	–	–	√	–	–	–	音频	图像	√	–	256×256
	AVCL(Wang 等,2021b)	√	–	√	–	–	–	√	–	–	–	音频	图像	√	–	256×256

注:“√”表示有此项内容,“–”表示无此项内容,“*”表示该方法仅修改源视频的人脸区域,即输出分辨率与输入源视频分辨率保持一致。

题定义尚不规范。这些方面均值得进一步研究。

4　评价指标

面部重演模型的评价指标通常从图像质量和人脸特征相似程度两方面对图像的生成质量以及人脸身份、表情和姿态等生成的准确程度进行评估。

基于图像生成质量的评价指标常常复用经典的图像生成的评价指标,包括峰值信噪比(PSNR)、结构相似性(SSIM)(Wang 等,2004)、累计模糊检测概率(CPBD)(Narvekar 和 Karam,2009)、FID(frechet inception distance)(Heusel 等,2018)等。用户评价(user study,US)亦用于评估图像生成质量。US 指标采取不同的算法使用相同的人脸图像数据进行重演,受试者会对不同的重演结果给出偏好信息。

此外,根据面部重演任务的特性,提出了一些针对人脸特征属性的评价指标。面部重演生成的人脸图像保留了源图像的身份信息和驱动图像的姿态、表情等动作信息,Ha 等人(2020)首次在面部重演任务中使用了 3 种从上述语义层面进行评估的指标,分别为人脸身份特征的余弦相似度(CSIM)、脸部转动角度的均方根误差(PRMSE)、面部动作单元的识别比例(AUCON)。

CSIM 对重演图像和提供身份信息的源图像分别使用人脸识别网络 Vggface2(Cao 等,2018)提取人脸特征向量,使用向量间的余弦相似度作为评价指标。计算为

$$f_{\mathrm{CSIM}} = \frac{\boldsymbol{f}_{\mathrm{r}} \cdot \boldsymbol{f}_{\mathrm{s}}}{\|\boldsymbol{f}_{\mathrm{r}}\| \|\boldsymbol{f}_{\mathrm{s}}\|} \tag{9}$$

式中,$\boldsymbol{f}_{\mathrm{r}}$ 和 $\boldsymbol{f}_{\mathrm{s}}$ 分别为重演图像和源图像的身份特征。

PRMSE 比较了重演图像与提供动作信息的驱动图像之间的人脸姿态角的差异程度,计算为

$$f_{\mathrm{PRMSE}} = \sqrt{\frac{\sum_{i=1}^{N} (\boldsymbol{A}_i - \hat{\boldsymbol{A}}_i)^2}{N}} \tag{10}$$

式中,N 为测试集总数,$\boldsymbol{A}$ 为真实人脸图像的姿态欧拉角,$\hat{\boldsymbol{A}}$ 则对应于重演图像。

AUCON 评估了重演图像与驱动图像之间面部动作单元的差异,计算为

$$f_{\mathrm{AUCON}} = \frac{1}{N} \sum_{i=1}^{N} \frac{\boldsymbol{AU}_i^{\mathrm{rec}}}{\boldsymbol{AU}_i^{\mathrm{d}}} \tag{11}$$

式中,$\boldsymbol{AU}^{\mathrm{d}}$ 为驱动人脸图像检测到的面部动作单元的总数,$\boldsymbol{AU}^{\mathrm{rec}}$ 为重演图像中识别出来且运动幅值相同的面部动作单元的总数。上述面部欧拉角和面部动作单元均由 OpenFace 工具(Baltrusaitis 等,2018)检测得到。

人脸关键点距离(LMD)也常用于评估重演图像生成的准确程度,计算为

$$f_{\mathrm{LMD}}=\sqrt{\sum_{i=1}^{n}(\boldsymbol{L}_i-\hat{\boldsymbol{L}}_i)^2} \tag{12}$$

式中,$\boldsymbol{L}$ 为真实人脸关键点坐标,$\hat{\boldsymbol{L}}$ 为生成的重演图像人脸关键点坐标,n 为人脸关键点总数。

在面对跨模态驱动面部重演的场景时,音唇一致性是面部重演模型需要考虑的重点问题,常见的做法是使用预训练的音视频一致判别网络(Chung 和 Zisserman,2017a)预测音频与视频的偏移量以及匹配的置信度分数。在用户评价方面,通过用户从音唇一致性、头部运动自然程度和视频真实性 3 方面的打分评价面部重演模型的效果。

5　常用的数据集

面部重演模型常用的数据集包含视频数据集和图像数据集,根据背景、光照的多样性又可分为单一场景和复杂场景两种。

5.1　图像数据集

5.1.1　单一场景下的图像数据集

RaFD(Radboud faces database)数据集(Langner 等,2010)由荷兰拉德堡德大学于 2010 年建立,包含 8 040 幅图像,拍摄场景为实验室场景,背景为单一的灰色,数据集包含 67 个人物,每个人物包含生气、恶心、恐惧、开心、悲伤、惊讶、轻视和中立等 8 种表情,包含左、中和右 3 种眼球朝向,每张照片由 5 个不同方向的相机捕捉拍摄。

5.1.2　复杂场景下的图像数据集

Multi-PIE(multi pose, illumination and expression)数据集(Gross 等,2010)由美国卡内基梅隆大学于 2010 年提出,包含 750 000 幅图像,拍摄场景为实验室场景,数据集包含 337 个人物,每个人物包含了微笑、惊讶、斜视、恶心和尖叫等表情,每张照片由 13 个不同方向的相机捕捉拍摄。此外,Multi-PIE 模拟了真实世界中 18 种不同程度的光照。

5.2　视频数据集

5.2.1　单一场景下的视频数据集

1)DeeperForensics 数据集(Jiang 等,2020)由新加坡南洋理工大学和商汤科技于 2020 年建立,在实验室受控环境下拍摄,背景为单一的绿色幕布,包含 60 000 条视频共 17 600 000 帧图像,其中 50 000 条视频为真实视频,数据集的人物主体包含了 100 位覆盖不同肤色的演员,部分视频包含了跨越 180°的头部姿态以及 9 种不同的光照条件。

2)MEAD(multi-view emotional audio-visual dataset)数据集(Wang 等,2020a)由商汤科技于 2020 年建立,包含 281 400 个视频片段,在实验室受控环境下拍摄,背景为单一的绿色幕布,数据集的人物主体为 60 名演员,包含了 8 种表情、3 个表情的动作幅度以及 7 个不同的摄像机拍摄视角。除视频模态外,MEAD 同样采集了说话人的音频信息。

5.2.2　复杂场景下的视频数据集

1)LRW(lip reading in the wild)数据集(Chung 和 Zisserman,2017b)。由英国牛津大学视觉几何组于 2016 年发布,起初用于语音识别。LRW 可用于音频驱动的面部重演,包含了 500 个不同单词的视频片段,每个单词对应近 1 000 条视频。

2)VoxCeleb(large-scale speaker identification dataset)数据集(Nagrani 等,2017)。由英国牛津大学视觉几何组于 2017 年发布,起初用于说话人识别,包含 1 251 位名人的约 10 万条脸部说话视频片段,均从采访或访谈类节目中截取。VoxCeleb 包含了说话人的身份标签,可用于音频驱动的面部重演模型研究。

3)VoxCeleb2 数据集(Chung 等,2018)。由英国牛津大学视觉几何组于 2017 年发布,VoxCeleb2 扩充了 VoxCeleb 的数据规模,包含了 6 112 位名人约 100 万条视频片段。

4)FaceForensics++数据集(Rössler 等,2019)由德国慕尼黑工业大学视觉计算与人工智能实验室在 2019 年建立,包含从网络上收集的 1 000 条真实新闻或采访的说话人视频片段,以及使用面部重演方法(Thies 等,2016,2019)生成的视频共 2 000 条。

5)VideoForensicsHQ 数据集(Fox 等,2021)由德国马克斯普朗克信息学研究所于 2021 年发布,包含了 1 737 个说话人视频,其中伪造视频帧 326 973 个,真实视频帧 1 339 843 个,视频中的人物包含 8 种表情,背景为真实场景。该数据集伪造视频使用的方法为 DVP(deep video portraits)(Kim 等,2018)。

6)ForgeryNet 数据集(He 等,2021)由商汤科技于 2021 年建立,包含了 99 630 个真实视频和 121 617 个伪造视频,真实视频帧和伪造视频帧均达到了百万级别,共包含 5 400 个不同身份的人。

该数据集的伪造视频的一部分由3种不同的面部重演模型生成得到(Chen等,2019;Siarohin等,2019b;Zakharov等,2019)。

6 面部重演的实际应用与潜在危害

随着面部重演领域的发展和视频门户网站与手机短视频社交平台的兴起,越来越多的用户开始接触到面部重演视频,享受着技术带来的娱乐与便捷,同时用户们也不可避免地会受到面部重演技术的滥用所带来的负面影响。

6.1 实际应用

Prajwal等人(2019)通过面部重演修改本地化配音后的外语电影中人物的嘴型,使音频与嘴型保持一致,提升了配音电影的观感。Lee(2019)将历史人物的面部重演配音后放入博物馆讲解视频中,提升了该人物的真实感。Wang等人(2021c)通过面部重演进行人脸为主的会议视频流压缩,仅通过传输源图像和后续视频帧的姿态特征、表情特征等参数,在用户端通过重演生成会议画面,视频质量明显优于同级别压缩率的其他方法。Nagano等人(2018)使用3维模型进行面部重演,仅通过单幅人脸图像即可得到该人物的重演化身,应用于相关游戏的角色创建,负责脸部动作的刻画。Japhet(2021)应用MyHeritage为黑白照片上色并使静止的人物活动,以及银行线上客服、虚拟教授、虚拟主播和3维数字人等均使用了面部重演技术。

6.2 潜在危害

面部重演属于深度伪造中的一种,可以在保持人物身份信息的情况下修改姿态、嘴型等动作信息,具有潜在危害。在结合音频的情况下,可以使用面部重演技术生成指定人物说任意话语的视频以诋毁他人、散布虚假信息等。例如,通过视频聊天窗口扮演他人以获取其家人信任、生成负面有害内容等。目前互联网视频网站上出现的外国人演唱中国歌曲的音视频片段(Xuan,2021),存在由面部重演模型生成的情况。

7 结 语

7.1 当前发展总结

面部重演起初的驱动模态仅包含图像,随着应用的需求衍化以及跨模态研究领域的发展,驱动模态逐步向文字和音频转化,进一步衍生了跨模态面部重演。与此同时,面部重演模型完成了推断时从有身份限制到无身份限制的转化。

7.1.1 图像驱动面部重演发展总结

基于图像驱动的面部重演方法按驱动特征的表达方式分为基于人脸关键点的方法、基于3维模型的方法、基于运动场预测的方法和基于特征解耦的方法。

基于人脸关键点的方法在实验室固定背景的图像数据集上表现较好,但是由于人脸关键点本身粒度较粗且无法表征人脸图像的背景信息,基于人脸关键点的方法难以表征连续的视频级别的重演动作以及大幅度的头部姿态变换。

基于3维模型的方法使用3维人脸形变模型对姿态、表情和身份信息进行解耦,可以有效应对不同姿态人脸的生成,然而其同样未对人脸背景信息进行建模,目前有相关研究使用人脸分割将背景分割出来并采用后处理的方法贴合人脸背景。

基于运动场预测的方法有效解决了背景生成不准确的问题,并可以生成连续的视频级别的重演动作。不足之处在于生成的重演人脸会存在一定的结构信息的偏差,原因在于提取运动场的驱动信息时,没有将驱动人物本身包含的身份信息显示剥离,导致提取的驱动信息混合了多种身份特征,最终导致了人脸结构信息的偏差,产生了身份不匹配问题。

基于特征解耦的方法可以有效地将不同语义空间的信息分离,生成语义准确的重演人脸图像。但是该方法对于特征提取的可解释性较差,并且忽略了人脸中本身包含的结构信息,同样会造成生成重演人脸的结构信息的不准确。

7.1.2 跨模态面部重演发展总结

跨模态面部重演主要包含基于文字驱动的面部重演和基于音频驱动的面部重演。

基于文字驱动的面部重演尚处在兴起阶段,常见的类型为根据文字材料修改特定人的说话视频的嘴部动作,该方法可扩展性差,对输入数据要求高,生成数据多样性差,生成人脸的面部姿态动作受到了输入源人脸的限制。

基于音频驱动的面部重演蓬勃发展,在生成的图像或视频的质量上可以达到以假乱真的水平,不过由于音频模态和视频模态之间存在着天然的信息

不匹配的问题。人脸图像的眨眼信息、头部姿态和面部表情等的建模准确性存在一定的上升空间，主要解决思路为采用特定的建模方法寻找到音频驱动信号与视觉模态的对应关系。此外，人说相同的语音内容时头部的微小动作也会是不同的，目前的跨模态面部重演仅局限在一对一的映射关系上，限制了重演模型生成的多样性。

7.2 未来发展方向

面部重演模型在模型优化和生成场景鲁棒性方面均存在潜在的发展方向。就模型优化而言，降低海量数据的强依赖、提升人物身份的泛化性、数据集的构建以及评价指标的建立均是未来可研究的方向。就生成场景鲁棒性而言，遮挡人脸、大角度人脸和真实场景下的人脸也是未来极具研究潜力的方向。

1)模型优化。面部重演模型很大程度上反映了训练数据的模式和分布。为了提高泛化性，研究者常用小样本学习或元学习的方式构建模型，不过这种方式需要海量的数据，使用超大规模的计算资源进行模型训练，如何提升泛化性的同时降低对数据的强依赖具有较大的研究挑战性。此外，目前包含重演视频的数据集较少，尚没有成对的可以直接应用于训练的数据集或专门为该任务准备的数据集，在语言方面尚无中文的说话人重演数据集以供训练。最后，面部重演的评价指标大多复用图像生成的评价指标，无法准确衡量帧间时序一致性和人们对图像中身份和表情变化的直观感受，无法直观评估人脸和背景衔接处的伪影严重程度。上述领域均值得进一步探索。

2)生成场景鲁棒性。人脸的遮挡、变换复杂的场景、大角度人脸均为生活中常见的场景，基于上述场景的面部重演也是未来发展的方向。遮挡指源人脸或驱动人脸被手、眼镜、头发和饰品等物体遮挡，由于说话人脸饰品数据集的限制，现有的面部重演模型几乎不考虑人脸遮挡问题。此外，头部姿态转动亦会导致背景信息的遮挡，常用的算法是采取遮挡图进行图像修复，不过效果较差。此外，变换复杂的场景包含了遮挡、光照和反射等一系列影响人脸表征的因素，然而大部分面部重演模型使用的是实验室内拍摄的人脸数据，这种人脸数据背景单一，无法反映真实场景下的说话人状态，影响了模型的广泛应用。最后，目前面部重演大多集中在小幅度的人脸区域的变动，无法模拟人头部的多种变化状态，在面部重演大角度人脸的时候，由于遮挡、背景等原因，重演图像的质量和效果会有很大下降。面对真实场景的面部重演生成有助于技术落地，值得进一步探索。

参考文献(References)

Averbuch-Elor H, Cohen-Or D, Kopf J and Cohen M F. 2017. Bringing portraits to life. ACM Transactions on Graphics, 36(6): #196 [DOI: 10.1145/3130800.3130818]

Baltrusaitis T, Zadeh A, Lim Y C and Morency L P. 2018. OpenFace 2.0: facial behavior analysis toolkit//Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition. Xi′an, China: IEEE: 59-66 [DOI: 10.1109/FG.2018.00019]

Blanz V and Vetter T. 1999. A morphable model for the synthesis of 3D faces//Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. New York, USA: ACM: 187-194 [DOI: 10.1145/311535.311556]

Burkov E, Pasechnik I, Grigorev A and Lempitsky V. 2020. Neural head reenactment with latent pose descriptors//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 13783-13792 [DOI: 10.1109/CVPR42600.2020.01380]

Cao Q, Shen L, Xie W D, Parkhi O M and Zisserman A. 2018. VGGFace2: a dataset for recognising faces across pose and age//Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition. Xi′an, China: IEEE: 67-74 [DOI: 10.1109/FG.2018.00020]

Chen L L, Li Z H, Maddox R K, Duan Z Y and Xu C L. 2018. Lip movements generation at a glance//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 538-553 [DOI: 10.1007/978-3-030-01234-2_32]

Chen L L, Maddox R K, Duan Z Y and Xu C L. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7824-7833 [DOI: 10.1109/CVPR.2019.00802]

Chung J S, Jamaludin A and Zisserman A. 2017. You said that? [EB/OL]. [2021-05-24]. https://arxiv.org/pdf/1705.02966.pdf

Chung J S, Nagrani A and Zisserman A. 2018. VoxCeleb2: deep speaker recognition [EB/OL]. [2021-05-02]. https://arxiv.org/pdf/1806.05622.pdf

Chung J S and Zisserman A. 2017a. Out of time: automated lip sync in the wild//Proceedings of 2016 ACCV International Workshops on Computer Vision. Taipei, China: Springer: 251-263 [DOI: 10.1007/978-3-319-54427-4_19]

Chung J S and Zisserman A. 2017b. Lip reading in the wild//Proceedings of the 13th Asian Conference on Computer Vision. Taipei, China: Springer: 87-103 [DOI: 10.1007/978-3-319-54184-6_6]

Fox G, Liu W T, Kim H, Seidel H P, Elgharib M and Theobalt C. 2021. Videoforensicshq: detecting high-quality manipulated face videos [EB/OL]. [2021-05-24]. https://arxiv.org/pdf/2005.10360.pdf

Fried O, Tewari A, Zollhöfer M, Finkelstein A, Shechtman E, Goldman D B, Genova K, Jin Z Y, Theobalt C and Agrawala M. 2019. Text-based editing of talking-head video. ACM Transactions on Graphics, 38(4): #68 [DOI: 10.1145/3306346.3323028]

Fu C Y, Hu Y B, Wu X, Wang G L, Zhang Q and He R. 2021. High-fidelity face manipulation with extreme poses and expressions. IEEE Transactions on Information Forensics and Security, 16: 2218-2231 [DOI: 10.1109/TIFS.2021.3050065]

Geng J H, Shao T J, Zheng Y Y, Weng Y L and Zhou K. 2018. Warp-guided GANs for single-photo facial animation. ACM Transactions on Graphics, 37(6): #231 [DOI: 10.1145/3272127.3275043]

Gross R, Matthews I, Cohn J, Kanade T and Baker S. 2010. Multi-PIE. Image and Vision Computing, 28(5): 807-813 [DOI: 10.1016/j.imavis.2009.08.002]

Gu K X, Zhou Y Q and Huang T. 2020. FLNet: landmark driven fetching and learning network for faithful talking facial animation synthesis. Proceedings of the AAAI Conference on Artificial Intelligence, 34(7): 10861-10868 [DOI: 10.1609/aaai.v34i07.6717]

Ha S, Kersner M, Kim B, Seo S and Kim D. 2020. MarioNETte: few-shot face reenactment preserving identity of unseen targets. Proceedings of the AAAI Conference on Artificial Intelligence, 34(7): 10893-10900 [DOI: 10.1609/aaai.v34i07.6721]

He Y, Gan B, Chen S Y, Zhou Y C, Yin G J, Song L C, Sheng L, Shao J and Liu Z W. 2021. ForgeryNet: a versatile benchmark for comprehensive forgery analysis [EB/OL]. [2021-05-24]. https://arxiv.org/pdf/2103.05630.pdf

Heusel M, Ramsauer H, Unterthiner T, Nessler B and Hochreiter S. 2018. GANs trained by a two time-scale update rule converge to a local nash equilibrium [EB/OL]. [2021-05-02]. https://arxiv.org/pdf/1706.08500.pdf

Huang P H, Yang F E and Wang Y C F. 2020. Learning identity-invariant motion representations for cross-ID face reenactment//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 7082-7090 [DOI: 10.1109/CVPR42600.2020.00711]

Isola P, Zhu J Y, Zhou T H and Efros A A. 2017. Image-to-image translation with conditional adversarial networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5967-5976 [DOI: 10.1109/CVPR.2017.632]

Jalalifar S A, Hasani H and Aghajan H. 2018. Speech-driven facial reenactment using conditional generative adversarial networks [EB/OL]. [2021-05-02]. https://arxiv.org/pdf/1803.07461.pdf

Japhet G. 2021. Animate your family photos [EB/OL]. [2021-05-26]. https://www.myheritage.tw/deep-nostalgia

Ji X Y, Zhou H, Wang K S Y, Wu W, Loy C C, Cao X and Xu F. 2021. Audio-driven emotional video portraits [EB/OL]. [2021-05-24]. https://arxiv.org/pdf/2104.07452.pdf

Jiang L M, Li R, Wu W, Qian C and Loy C C. 2020. Deeperforensics-1.0: a large-scale dataset for real-world face forgery detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 2886-2895 [DOI: 10.1109/CVPR42600.2020.00296]

Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J and Aila T. 2020. Analyzing and improving the image quality of StyleGAN//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8107-8116 [DOI: 10.1109/CVPR42600.2020.00813]

Kim H, Garrido P, Tewari A, Xu W P, Thies J, Niessner M, Pérez P, Richardt C, Zollhöfer M and Theobalt C. 2018. Deep video portraits. ACM Transactions on Graphics, 37(4): #163 [DOI: 10.1145/3197517.3201283]

Langner O, Dotsch R, Bijlstra G, Wigboldus D H J, Hawk S T and Van Knippenberg A. 2010. Presentation and validation of the Radboud Faces Database. Cognition and Emotion, 24(8): 1377-1388 [DOI: 10.1080/02699930903485076]

Lee D. 2019. Deepfake Salvador Dalí takes selfies with museum visitors [EB/OL]. [2021-05-02]. https://www.theverge.com/2019/5/10/18540953/salvador-dali-lives-deepfake-museum

Li L C, Wang S Z, Zhang Z M, Ding Y, Zheng Y X, Yu X and Fan C J. 2021a. Write-a-speaker: text-based emotional and rhythmic talking-head generation. Proceedings of the AAAI Conference on Artificial Intelligence, 35(3): 1911-1920

Li X R, Ji S L, Wu C M, Liu Z G, Deng S G, Cheng P, Yang M and Kong X W. 2021b. Survey on deepfakes and detection techniques. Journal of Software, 32(2): 496-518 (李旭嵘, 纪守领, 吴春明, 刘振广, 邓水光, 程鹏, 杨珉, 孔祥维. 2021b. 深度伪造与检测技术综述. 软件学报, 32(2): 496-518) [DOI: 10.13328/j.cnki.jos.006140]

Liu J, Chen P, Liang T, Li Z X, Yu C, Zou S Q, Dai J and Han J Z. 2021. LI-Net: large-pose identity-preserving face reenactment network [EB/OL]. [2021-05-07]. https://arxiv.org/pdf/2104.02850.pdf

Liu Z X, Hu H, Wang Z P, Wang K, Bai J Q and Lian S G. 2019. Video synthesis of human upper body with realistic face//2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct). Beijing, China: IEEE: 200-202 [DOI: 10.1109/ISMAR-Adjunct.2019.00-47]

Lyu S. 2020. Deepfake detection: current challenges and next steps//Proceedings of 2020 IEEE International Conference on Multimedia and Expo Workshops (ICMEW). London, UK: IEEE: 1-6 [DOI:

10.1109/ICMEW46912.2020.9105991]

Ma T X, Peng B, Wang W and Dong J. 2019. Any-to-one face reenactment based on conditional generative adversarial network//Proceedings of 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Lanzhou, China: IEEE: 1657-1664 [DOI: 10.1109/APSIPAASC47483.2019.9023328]

Martinez B, Valstar M F, Jiang B H and Pantic M. 2019. Automatic analysis of facial actions: a survey. IEEE Transactions on Affective Computing, 10(3): 325-347 [DOI: 10.1109/TAFFC.2017.2731763]

Mirsky Y and Lee W. 2022. The creation and detection of deepfakes: a survey. ACM Computing Surveys, 54(1): #7 [DOI: 10.1145/3425780]

Nagano K, Seo J, Xing J, Wei L Y, Li Z M, Saito S, Agarwal A, Fursund J and Li H. 2018. paGAN: real-time avatars using dynamic textures. ACM Transactions on Graphics, 37(6): #528 [DOI: 10.1145/3272127.3275075]

Nagrani A, Chung J S and Zisserman A. 2017. VoxCeleb: a large-scale speaker identification dataset [EB/OL]. [2021-05-02]. https://arxiv.org/pdf/1706.08612.pdf

Narvekar N D and Karam L J. 2009. A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection//Proceedings of 2009 International Workshop on Quality of Multimedia Experience. San Diego, USA: IEEE: 87-91 [DOI: 10.1109/QOMEX.2009.5246972]

Nirkin Y, Keller Y and Hassner T. 2019. FSGAN: subject agnostic face swapping and reenactment//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 7183-7192 [DOI: 10.1109/ICCV.2019.00728]

Prajwal K R, Mukhopadhyay R, Namboodiri V P and Jawahar C V. 2020. A lip sync expert is all you need for speech to lip generation in the wild//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 484-492 [DOI: 10.1145/3394171.3413532]

Prajwal KR, Mukhopadhyay R, Philip J, Jha A, Namboodiri V and Jawahar CV. 2019. Towards automatic face-to-face translation//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: ACM: 1428-1436 [DOI: 10.1145/3343031.3351066]

Pumarola A, Agudo A, Martinez A M, Sanfeliu A and Moreno-Noguer F. 2018. GANimation: anatomically-aware facial animation from a single image//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 835-851 [DOI: 10.1007/978-3-030-01249-6_50]

Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J and Nießner M. 2019. FaceForensics++: learning to detect manipulated facial images//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 1-11 [DOI: 10.1109/ICCV.2019.00009]

Sanchez E and Valstar M. 2018. Triple consistency loss for pairing distributions in GAN-based face synthesis [EB/OL]. [2021-05-24]. https://arxiv.org/pdf/1811.03492.pdf

Shen Y J, Luo P, Luo P, Yan J J, Wang X G and Tang X O. 2018a. FaceID-GAN: learning a symmetry three-player GAN for identity-preserving face synthesis//Proceedings of 2018 IEEE/CVF conference on computer vision and pattern recognition. Salt Lake City, USA: IEEE: 821-830 [DOI: 10.1109/CVPR.2018.00092]

Shen Y J, Zhou B L, Luo P and Tang X O. 2018b. FaceFeat-GAN: a two-stage approach for identity-preserving face synthesis [EB/OL]. [2021-05-25]. https://arxiv.org/pdf/1812.01288.pdf

Siarohin A, Lathuilière S, Tulyakov S, Ricci E and Sebe N. 2019a. Animating arbitrary objects via deep motion transfer//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 2372-2381 [DOI: 10.1109/CVPR.2019.00248]

Siarohin A, Lathuilière S, Tulyakov S, Ricci E and Sebe N. 2019b. First order motion model for image animation//Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 7137-7147

Song L S, Wu W, Qian C, He R and Loy C C. 2022. Everybody's talkin': let me talk as you want. IEEE Transactions on Information Forensics and Security, 17: 585-598 [DOI: 10.1109/TIFS.2022.3146783]

Song Y, Zhu J W, Li D W, Wang X and Qi H R. 2019. Talking face generation by conditional recurrent adversarial network [EB/OL]. [2021-12-29]. https://arxiv.org/pdf/1804.04786.pdf

Sun P, Li Y Z, Qi H G and Lyu S W. 2021. LandmarkGAN: synthesizing Faces from Landmarks [EB/OL]. [2021-05-25]. https://arxiv.org/pdf/2011.00269.pdf

Suwajanakorn S, Seitz S M and Kemelmacher-Shlizerman I. 2017. Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics, 36(4): #95 [DOI: 10.1145/3072959.3073640]

Thies J, Elgharib M, Tewari A, Theobalt C and Nießner M. 2020. Neural voice puppetry: audio-driven facial reenactment//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 716-731 [DOI: 10.1007/978-3-030-58517-4_42]

Thies J, Zollhöfer M and Nießner M. 2019. Deferred neural rendering: Image synthesis using neural textures. ACM Transactions on Graphics, 38(4): #66 [DOI: 10.1145/3306346.3323035]

Thies J, Zollhöfer M, Nießner M, Valgaerts L, Stamminger M and Theobalt C. 2015. Real-time expression transfer for facial reenactment. ACM Transactions on Graphics, 34(6): #183 [DOI: 10.1145/2816795.2818056]

Thies J, Zollhöfer M, Stamminger M, Theobalt C and Nießner M. 2016. Face2face: real-time face capture and reenactment of RGB videos//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 2387-2395 [DOI: 10.1109/CVPR.2016.262]

Tolosana R, Romero-Tapiador S, Fierrez J and Vera-Rodriguez R. 2020a. DeepFakes evolution: analysis of facial regions and fake detection performance [EB/OL]. [2021-05-27]. https://arxiv.org/pdf/2004.07532.pdf

Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A and Ortega-Garcia J. 2020b. Deepfakes and beyond: a survey of face manipulation and fake detection. Information Fusion, 64: 131-148 [DOI: 10.1016/j.inffus.2020.06.014]

Tripathy S, Kannala J and Rahtu E. 2021. FACEGAN: facial attribute controllable rEenactment GAN//Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 1328-1337 [DOI: 10.1109/WACV48630.2021.00137]

Wang K S Y, Wu Q Y, Song L S, Yang Z Q, Wu W, Qian C, He R, Qiao Y and Loy C C. 2020a. MEAD: a large-scale audio-visual dataset for emotional talking-face generation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 700-717 [DOI: 10.1007/978-3-030-58589-1_42]

Wang S Z, Li L C, Ding Y, Fan C J and Yu X. 2021a. Audio2Head: audio-driven one-shot talking-head generation with natural head motion [EB/OL]. [2022-02-17]. https://arxiv.org/pdf/2107.09293.pdf

Wang S Z, Li L C, Ding Y and Yu X. 2021b. One-shot talking face generation from single-speaker audio-visual correlation learning [EB/OL]. [2022-06-22]. https://arxiv.org/pdf/2112.02749.pdf

Wang T C, Liu M Y, Zhu J Y, Liu G L, Tao A, Kautz J and Catanzaro B. 2018. Video-to-video synthesis [EB/OL]. [2022-02-17]. https://arxiv.org/pdf/1808.06601.pdf

Wang T C, Mallya A and Liu M Y. 2021c. One-shot free-view neural talking-head synthesis for video conferencing [EB/OL]. [2021-05-02]. https://arxiv.org/pdf/2011.15126.pdf

Wang Y H, Bilinski P, Bremond F and Dantcheva A. 2020b. ImaGINator: conditional spatio-temporal GAN for video generation//Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision. Snowmass, USA: IEEE: 1149-1158 [DOI: 10.1109/WACV45572.2020.9093492]

Wang Z, Bovik A C, Sheikh H R and Simoncelli E P. 2004. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4): 600-612 [DOI: 10.1109/TIP.2003.819861]

Wiles O, Koepke A S and Zisserman A. 2018. X2Face: a network for controlling face generation using images, audio, and pose codes//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 690-706 [DOI: 10.1007/978-3-030-01261-8_41]

Wu W, Zhang Y X, Li C, Qian C and Loy C C. 2018. ReenactGAN: learning to reenact faces via boundary transfer//Proceedings of the 15th European Conference on Computer Vision (ECCV) Munich, Germany: Springer: 622-638 [DOI: 10.1007/978-3-030-01246-5_37]

Xiang S T, Gu Y M, Xiang P D, He M M, Nagano K, Chen H W and Li H. 2020. One-shot identity-preserving portrait reenactment [EB/OL]. [2021-05-26]. https://arxiv.org/pdf/2004.12452.pdf

Xuan S X. 2021. Video of Donald Trump singing [EB/OL]. [2021-05-26]. https://www.bilibili.com/video/BV1Xz4y1U7EY

Yao G M, Yuan Y, Shao T J, Li S, Liu S Q, Liu Y, Wang M M and Zhou K. 2021. One-shot face reenactment using appearance adaptive normalization [EB/OL]. [2021-05-07]. https://arxiv.org/pdf/2102.03984.pdf

Yao G M, Yuan Y, Shao T J and Zhou K. 2020a. Mesh guided one-shot face reenactment using graph convolutional networks//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 1773-1781 [DOI: 10.1145/3394171.3413865]

Yao X W, Fried O, Fatahalian K and Agrawala M. 2020b. Iterative text-based editing of talking-heads using neural retargeting [EB/OL]. [2021-05-24]. https://arxiv.org/pdf/2011.10688.pdf

Yu L Y, Yu J and Ling Q. 2019. Mining audio, text and visual information for talking face generation//Proceedings of 2019 IEEE International Conference on Data Mining (ICDM). Beijing, China: IEEE: 787-795 [DOI: 10.1109/ICDM.2019.00089]

Zakharov E, Ivakhnenko A, Shysheya A and Lempitsky V. 2020. Fast Bi-layer neural synthesis of one-shot realistic head avatars//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 524-540 [DOI: 10.1007/978-3-030-58610-2_31]

Zakharov E, Shysheya A, Burkov E and Lempitsky V. 2019. Few-shot adversarial learning of realistic neural talking head models//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 9458-9467 [DOI: 10.1109/ICCV.2019.00955]

Zeng X F, Pan Y S, Wang M M, Zhang J N and Liu Y. 2020. Realistic face reenactment via self-supervised disentangling of identity and pose. Proceedings of the AAAI Conference on Artificial Intelligence, 34(7): 12757-12764 [DOI: 10.1609/aaai.v34i07.6970]

Zhang J N, Liu L, Xue Z C and Liu Y. 2020a. APB2Face: audio-guided face reenactment with auxiliary pose and blink signals//Proceedings of ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE: 4402-4406 [DOI: 10.1109/ICASSP40776.2020.9052977]

Zhang J N, Zeng X F, Wang M M, Pan Y S, Liu L, Liu Y, Ding Y and Fan C J. 2020b. FreeNet: multi-identity face reenactment//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5325-5334 [DOI: 10.1109/CVPR42600.2020.00537]

Zhang J N, Zeng X F, Xu C, Chen J, Liu Y and Jiang Y L. 2020c. APB2FaceV2: real-time audio-guided multi-face reenactment [EB/OL]. [2021-05-24]. https://arxiv.org/pdf/2010.13017.pdf

Zhang Y X, Zhang S W, He Y, Li C, Loy C C and Liu Z W. 2019. One-shot face reenactment [EB/OL]. [2021-05-26]. https://arxiv.org/pdf/1908.03251.pdf

Zhang Z M, Li L C, Ding Y and Fan C J. 2021. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset// Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 3660-3669 [DOI: 10.1109/CVPR46437.2021.00366]

Zhou H, Liu Y, Liu Z W, Luo P and Wang X G. 2019. Talking face generation by adversarially disentangled audio-visual representation. Proceedings of the AAAI Conference on Artificial Intelligence, 33(1): 9299-9306 [DOI: 10.1609/aaai.v33i01.33019299]

Zhou H, Sun Y S, Wu W, Loy C C, Wang X G and Liu Z W. 2021. Pose-controllable talking face generation by implicitly modularized audio-visual representation [EB/OL]. [2021-05-07]. https://arxiv.org/pdf/2104.11116.pdf

Zhou Y, Han X T, Shechtman E, Echevarria J, Kalogerakis E and Li D. 2020. MakeItTalk: speaker-aware talking-head animation. ACM Transactions on Graphics, 39(6): #221 [DOI: 10.1145/3414685.3417774]

作者简介

刘锦，1996 年生，男，博士研究生，主要研究方向为计算机视觉、图像生成、深度伪造的生成和检测。
E-mail：liujin@iie.ac.cn

王茜，通信作者，女，助理研究员，主要研究方向为多媒体智能信息处理、计算机视觉、多模态虚假信息检测。
E-mail：wangxi1@iie.ac.cn

陈鹏，男，博士研究生，主要研究方向为计算机视觉、人工智能安全、目标检测、对抗样本以及深度伪造的生成和检测。
E-mail：chenpeng@iie.ac.cn

付晓蒙，男，硕士研究生，主要研究方向为计算机视觉、图像合成。E-mail：fuxiaomeng@iie.ac.cn

戴娇，女，高级工程师，主要研究方向为多媒体信息智能化处理、多媒体内容理解和人工智能安全。
E-mail：daijiao@iie.ac.cn

韩冀中，男，正高级工程师，主要研究方向为大数据存储与管理、多媒体信息智能化处理、多媒体内容理解、深度伪造。
E-mail：hanjizhong@iie.ac.cn