ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 5,Sep. 2021,pp. 898－920
DOI:10. 16337/j. 1004-9037. 2021. 05. 006

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

人脸表情合成算法综述

郭迎春，王静洁，刘　依，夏伟毅，张吉俊，李学博，王天瑞

（河北工业大学人工智能与数据科学学院，天津 300400）

摘　要：人脸表情合成技术旨在保留人脸身份信息的情况下，对人脸表情进行重建，从而生成具有新表情的源人脸图像。深度学习的发展为表情合成提供了全新的解决方案，本文从特征提取、生成对抗网络的表情合成和实验评估方面综述了人脸表情合成技术的发展。首先，介绍了人脸特征的提取，这是表情合成任务中的一项关键技术，人脸特征可客观全面地描述人脸表情状态。其次，分析了表情合成领域中主流的基于深度学习的方法，主要针对生成对抗网络（Generative adversarial network，GAN）的发展现状，探讨了基于生成对抗网络的表情合成方法。通过对人脸数据集及实验评估方法的深入研究，总结出广泛使用的人脸表情合成数据集以及多种客观评价方法。最后根据现有方法所存在的问题，提出了未来工作的研究方向。

关键词：表情合成；深度学习；生成对抗网络；表情数据库；客观评价方法

中图分类号：TP391.4　　**文献标志码：**A

Survey on Facial Expression Synthesis Algorithms

GUO Yingchun, WANG Jingjie, LIU Yi, XIA Weiyi, ZHANG Jijun, LI Xuebo, WANG Tianrui

(School of Artificial Intelligence, Hebei University of Technology, Tianjin 300400, China)

Abstract: Facial expression synthesis technology is designed to reconstruct face image with new expressions while retaining identity information. The development of deep learning provides a new solution for the synthesis of facial expressions. This paper introduces the development of facial expression synthesis technology from the aspects of feature extraction, expression synthesis of generated antagonistic networks and experimental evaluation. Firstly, extraction of facial features is introduced, which is the key technology in expression synthesis. Facial features can describe facial expressions objectively and comprehensively. Secondly, the state-of-the-art facial expression synthesis methods based on deep learning are analyzed, in which methods based on generative adversarial network (GAN) are mainly discussed. By research on facial expression datasets and evaluation methods, the widely used facial expression datasets and objective evaluation methods are given in this paper. Finally, future work is discussed according to the existing problems of facial expression synthesis methods.

Key words: expression synthesis; deep learning; generative adversarial network; expression dataset; objective assessment methods

---

基金项目：国家自然科学基金（60302018，61806071）资助项目；河北省自然科学基金（F2019202381，F2019202464）资助项目。
收稿日期：2020-09-08；**修订日期：**2021-01-09

引　言

最近,“AI换脸”频频登上热搜,其主要技术就是表情合成,这是计算机图形学角色动画领域的关键技术之一。表情合成是一个经典的图形学问题,目的是在一个特定的目标人像上合成具有特定表情的人脸图像。随着科技的发展,如果将“AI换脸”技术应用在影视剧制作领域,可以大大减轻剪辑人员的工作量[1]。表情合成涉及计算机图形学、计算机视觉、图像处理、人机交互和面部解剖学等多个领域,可以应用在数字娱乐领域、人机交互、仿生代理及数据扩充等诸多领域[2]。因此,如何鲁棒地合成自然逼真的人脸表情成为一个富有挑战性的热点研究课题[3]。

人脸表情合成图像应当具有3个特点:(1)完成表情迁移的任务,即合成指定的目标表情,改变源图像中的人脸表情;(2)保留源人物的身份信息,即改变源人物表情的同时,不能改变源人物的身份特征;(3)合成图像应当具有合理、自然、逼真的面部细节,不应存在大量伪影或图像伪造痕迹。目前已提出的人脸表情合成方法大致可分为两大类:基于计算机图形学技术的传统方法和基于深度学习的表情合成方法。基于计算机图形学技术的传统方法主要研究目标合成图像与真实输入图像的人脸纹理之间的对应关系。早期方法[4-8]通过3D人脸纹理模型、特征对应或光流图等信息来扭曲人脸图像,或利用人脸补丁合成目标图像等。常用方法有表情映射[9]、图像变形[10]、图像检索[11-12]和统计学模型等。这些传统方法可以合成逼真的高分辨率的人脸图像。

近年来深度学习在很多领域都取得了突破性进展,亦为表情合成提供了一个全新的思路。在表情合成领域中,基于深度学习的方法首先利用卷积神经网络提取人脸特征,将图像从高维空间映射到低维的特征空间中,然后向网络中添加目标表情信息或更改原来的表情特征,再利用深度神经网络合成具有目标表情的人脸图像映射回高维空间。其中,目标表情信息属于高级语义信息,因此深度方法可以通过高级语义来控制图像的合成。

在表情合成方法的研究历程中,自1972年Parke[13]用计算机合成人脸图像之后的40多年内,基于计算机图形学技术的传统方法在这一领域占据主导地位。传统方法虽然可以合成清晰的人脸图像,但很难合成从未见过的人脸或人脸区域,如口内牙齿区域或皮肤皱纹,且传统方法更依赖于人工精心设计的复杂模型,每个模型都是针对具体任务设计的,泛化能力及鲁棒性较差。与之相比,深度方法是数据驱动的、基于大量训练样本的特征学习,所提取的抽象特征鲁棒性更强,泛化能力更好。但深度方法目前存在的问题是很难对图像的合成进行细粒度控制,且合成的表情图像缺乏人脸细节的表达,如皱纹或由于凹陷产生的阴影现象等。随着深度学习的发展,2014年Goodfellow等[14]提出了生成对抗网络(Generative adversarial network,GAN)。GAN作为一种近年来生成效果较好的模型,在合成逼真图像方面具有显著优势。同时,表情合成领域也涌现出大量的GAN变体等研究方法,基于GAN的表情合成方法目前已成为本领域的主流方法。

本文综述了人脸特征信息的分类、基于深度学习的人脸表情生成的3大模型以及主要的损失函数;着重介绍现有的人脸表情生成数据库;总结了表情生成的实验评估方法以及现有研究的评价方法;对表情生成方法存在的问题进行了分析讨论,并对未来的研究方向和发展趋势做出了展望。

1　人脸特征信息

在表情合成领域,包含人脸特征的标签信息对表情合成的质量和效果至关重要。尤其在深度学习方法中,标签作为监督信号输入,引导网络的训练过程,其作用不容忽视。一般最常见的标签是表情类别标签。为了使标签包含更多的人脸信息,研究人员使用人脸几何特征作为标签。这些几何特征标签有动作单元(Action unit, AU)、人脸关键点和人脸轮廓图等。

1.1 表情类别标签

表情类别标签是最基本的情绪标签，也是表情数据库中最常见的一种标签。一般将人的基本表情分为8类，分别是：快乐、悲伤、惊讶、愤怒、蔑视、厌恶、恐惧和中性。表情类别标签的获取，可以从数据库中直接获得；随着表情识别技术的发展，研究人员提出了很多可靠的表情识别算法，可以通过这些算法来获得表情类别标签。传统的表情识别算法首先通过主成分分析、独立分量分析、Gabor小波、局部二值模式(Local binary pattern，LBP)算子和光流等方法进行表情特征提取，然后使用 K 最近邻(K-nearest neighbor，KNN)、隐马尔可夫模型(Hidden Markov model，HMM)、贝叶斯分类算法和支持向量机(Support vector machines，SVM)等传统机器学习算法，根据提取到的特征将输入图片分为某种基本表情类别。随着深度学习的发展，很多方法应用在表情识别领域，端到端地输出表情类别。在卷积神经网络中，研究人员一般在网络的末端使用softmax分类器或SVM等传统分类算法来预测输入图片的表情类别。随着表情识别算法发展得越来越成熟，很多平台向用户提供了识别 应用程序接口(Application programming interface，API)，这些算法首先检测出图片中的人脸区域，然后针对脸部的表情返回识别结果。

1.2 面部动作编码系统

面部动作编码系统(Facial action coding system，FACS)是描述面部表情较为全面、客观的系统之一，近年来在情感计算和计算机视觉领域受到越来越多的关注。表情是脸部肌肉联合和协调作用的结果[15]，不能被划分为离散的或少数的类别。1978年，心理学家Ekman和Friesen开发了FACS系统，以AU来描述表情状态[16]。FACS系统描绘了脸部肌肉动作和表情之间的对应关系。根据人脸的解剖学特点，Ekman等[16]将人脸划分为若干个既相互独立又相互联系的AU，每个AU有6个强度，并分析了这些AU的运动特征及其所控制的主要区域及与之相关的表情。FACS中共有44个AU，直接与特定面部肌肉收缩有关的AU有30个，虽然AU的数量相对较少，但已经观察到有超过7 000种不同的AU组合。如图1所示，AU1代表抬起眉毛内角，AU2代表抬起眉毛外角，AU6代表脸颊提升和眼轮匝肌外圈收紧，AU9代表皱鼻肌，AU25、AU26、AU27代表嘴巴张开的程度等[17]。不同的AU组合可以表现不同的表情，如图2所示，AU9＋AU25可表现为“愤怒”，AU9＋AU16＋AU25表现为“非常愤怒”，AU6＋AU12表现为“微笑”，AU25代表张开嘴巴，因此AU6＋AU12＋AU25展示为“露着牙齿的笑”。

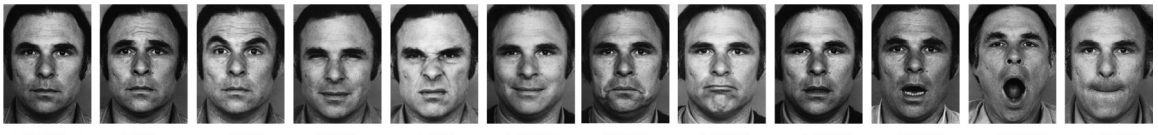


图1 动作单元示例

Fig.1 Example diagrams of AUs

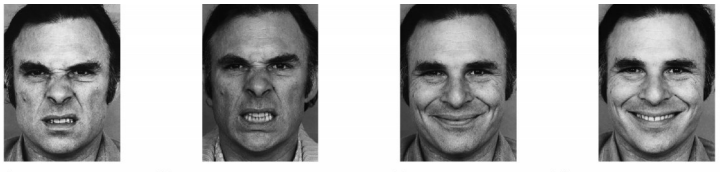


图2 动作单元组合示例

Fig.2 Example diagrams of combination of AUs

Liu等[18]将12个AU的6个强度线性缩放到[−1,1],每个AU有相应的强度值,形成一个12维的向量以表示目标表情。通过将此向量标签输入到模型中,指导网络的训练方向,生成带有目标表情的人脸图像。

1.3 人脸关键点

人脸关键点检测是人脸相关任务中一个关键步骤,如表情识别、人脸识别、表情合成和人脸验证等,借助它可以提取丰富的人脸特征信息和表情信息[19]。人脸关键点检测是对给定的人脸图像定位出人脸区域的关键位置,包括眉毛、眼睛、鼻子、嘴巴和脸部轮廓等,是表情合成任务中的基础环节,因此产生了很多检测方法[20-24],如何获取高精度人脸关键点,一直以来都是计算机视觉、模式识别和图像处理等领域的热点研究问题。

在标注关键点过程中,关键点数量的选择尤为重要。如果关键点数量过少,会导致人脸表情数据信息丢失过多,从而影响最终的人脸表情合成效果;相反,过多的关键点会导致标定难度增大,而且容易产生误差较大的离群点,同时也可能产生冗余信息,影响人脸表情合成的效果。目前多数表情合成方法采取68个关键点进行研究。图3为使用Dlib机器学习库对人脸图像标注面部关键点的示例,图中选择使用68个关键点,以白色圆点表示。68个关键点可包含大部分的人脸特征信息,提取的关键点包括了眉毛的形状、眼睛的轮廓、鼻子边框线、嘴巴的上下嘴唇的位置以及人脸的外轮廓等信息。

图3 人脸关键点示例

Fig.3 Example diagrams of facial landmarks

人脸关键点信息是衡量表情表达的一项重要指标,因此研究人员用人脸关键点来指导监督表情合成的过程。Song等[25]提出的G2-GAN(Geometry-guided generative adversarial network)网络中,将目标表情的人脸关键点热图和中性图像一起输入网络中,合成带有目标表情的人脸图像。G2-GAN中有2个生成器来执行相反的任务,表情合成和表情移除。在表情合成过程中,人脸关键点起到引导控制图像的合成的作用;在表情移除过程中,人脸关键点表示对输入图像的解释说明。Qiao等[26]提出的GC-GAN(Geometry-contrastive generative adversarial network)网络中,将目标表情关键点图像和输入人脸图像使用编码器压缩为特征向量,然后将这两个特征向量级联,输入到解码器中生成带有目标表情的人脸图像。

1.4 人脸轮廓图

在使用人脸几何信息指导表情合成过程中,除了人脸关键点信息,还可以将人脸关键点改进为人脸轮廓特征图来指导图像的合成。在标定人脸关键点后,将人脸关键点用线段连接,得到人脸轮廓特征图如图4所示。

图4 人脸轮廓图示例

Fig.4 Facial contour sketches

文献[27]将面试者的图像编码为特征图,与面试官的目标表情轮廓图融合输入到网络中,生成面试官的目标表情图像。文献[28]将局部缺损的输入图像和相应的局部轮廓图等信息一起输入模型中,生成补全的完整原始图像。文献[27-31]表明,人脸关键点或人脸轮廓图等这些几何标签可以对表情的合成起到一定的引导作用。

表1总结了人脸特征信息的获取方法。一般地,研究人员可选取相应的人脸数据库直接获得人脸特征信息,如表情类别、AU、人脸关键点坐标等。其次,研究人员可采用成熟的相关识别算法或检测算法来得到人脸特征信息,如手工特征方法或预训练的深度模型等。

表1 人脸特征信息获取方式

Table 1 Methods of extracting facial features

人脸特征信息	获取方法	代表文献/数据库
表情类别标签	数据库标注	CK+[①], MMI[②], JAFFE[③], TFD, FER-2013[④], Multi-PIE[⑤]
	表情识别手工特征算法	KNN[32], HMM[33], 贝叶斯分类算法[34], SVM[35]
	表情识别深度算法	CNN[36], DBN[37], DAE[38], RNN[39]
	在线API	Microsoft Azure, Baidu AI开放平台, 腾讯优图AI开放平台
动作单元AU	数据库标注	DISFA[⑥], BP4D[⑦], EmotioNet[⑧], CFEE[⑨], CK+
	AU识别手工特征算法	AU+GentleBoost[40], AU+Dictionary[41], Haar+Adaboost[42], JPML[43]
	AU识别深度学习算法	DRML[17], EAC-Net[44], JAA-Net[45], DSIN[46]
	专家标注	TCAE[47], AU R-CNN[48]
人脸关键点	数据库标注	IMM[⑩], FGnet[⑪], Helen[⑫], 300-W[⑬], IBUG[⑭]
	人脸开源库	Dlib[⑮], Face++[⑯], OpenCV[⑰]
	人脸关键点检测算法	ASM[20], AAM[22], CPR[23], DCNN[24], TCDCN[49], MTCNN[50], TCNN[51], DAN[52]
人脸轮廓图	基于关键点的连接	TCNN[51]
	基于曲线	DyadGAN[27]

①http://vasc.ri.cmu.edu/idb/html/face/facial_expression/.
②https://mmifacedb.eu/.
③http://www.kasrl.org/jaffe.html.
④https://www.kaggle.com/deadskull7/fer2013.
⑤http://www.multipie.org/.
⑥http://mohammadmahoor.com/disfa/.
⑦https://doi.org/10.1016/j.imavis.2014.06.002.
⑧https://doi.org/10.1109/CVPR.2016.600.
⑨https://www.pnas.org/content/111/15/E1454.full.
⑩http://www.imm.dtu.dk/pubdb/edoc/imm3160.pdf.
⑪https://doi.org/doi:10.18129/B9.bioc.FGNet.
⑫http://www.f-zhou.com/fa_code.html.
⑬https://ibug.doc.ic.ac.uk/resources/300-W/.
⑭https://ibug.doc.ic.ac.uk/resources/facial-point-annotations/.
⑮http://blog.dlib.net/.
⑯https://www.faceplusplus.com.cn/.
⑰https://docs.opencv.org/master/d2/d42/tutorial_face_landmark_detection_in_an_image.html.

2 深度学习方法

在图像合成领域,3大深度生成模型为自编码器(AutoEncoder,AE)、GAN及GAN变体。这3种深度模型同样在表情合成领域占据垄断地位。早期的深度方法使用AE的编解码结构来生成人脸,通过编码器和解码器来实现特征提取及生成图像等操作。由于AE算法的固有缺点,其偏向于生成模糊的图像,这限制了自编码器的发展。而后GAN网络的出现可以帮助合成清晰的图像,因此在图像合成领域涌现了大量的GAN变体等研究方法。

2.1 自编码器

自编码器包含编码器和解码器两部分,典型结构如图5所示。输入原图像,首先利用编码器将输入图像压缩为一个特征向量;然后通过解码器将特征向量重构回图像。学习过程中不断地最小化输出数据与输入数据之间的差异,从而达到图像重构的目的。

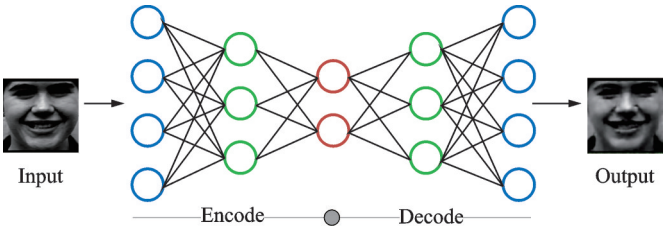


图5 自编码器的一般结构

Fig.5 General structure of AE

Zhou等[53]提出的条件差分对抗自编码器(Conditional difference adversarial autoencoder, CDAAE)是利用自编码器合成人脸表情的一个典型模型。该模型输入一幅带有任意表情的人脸图像,经编码器编码为一个100维的特征向量,然后将此向量与目标表情标签向量级联输入到解码器中,生成带有目标表情的人脸图像。Zhou等采用U-Net[54]的跳跃连接结构来保留人脸身份信息,并保证图像的合成质量。Zhou等还测试了在不同的网络层级间使用跳跃连接,可以不同程度地保留目标图像底层的纹理信息,并且使用自编码器结构合成图像后,采用对抗的思想加入2个判别器结构以保证合成的图像更为逼真。

自编码器的原理中,首先定义真实的样本分布为$\tilde{p}(x)$,然后构建一个带参数的后验分布$p(z|x)$,表示由z来生成x的模型,两者组成一个联合分布$p(x,z)=\tilde{p}(x)\,p(z|x)$。接着,定义一个先验分布$q(z)$和一个生成分布$q(x|z)$,这两者组成另一个联合分布$q(x,z)=q(z)\,q(x|z)$。自编码器的目的是使$p(x,z)$和$q(x,z)$互相逼近,优化二者之间的KL(Kullback-Leibler)散度$\mathrm{KL}(p(x,z)\|q(x,z))$。但在自编码器系统中,$p(z|x)$,$q(z)$,$q(x|z)$都被设为各分量独立的高斯分布,因此不能拟合任意复杂的分布,意味着$\mathrm{KL}(p(x,z)\|q(x,z))$从理论上来说不可能为0,所以$p(x,z)$和$q(x,z)$互相逼近的最终结果只能得到一个大致、平均的结果,这是自编码器生成图像偏模糊的原因。

2.2 生成对抗网络及其变体

GAN和自编码器的结构相似,都采用编解码器作为图像生成器。不同的是,GAN使用对抗博弈的思想对网络进行训练,通过引入一个判别器,形成生成器与判别器对抗的局面。判别器的目的是对输入的图像做出正确的真假预测,而生成器的目的是使生成图像足够逼真到骗过判别器,二者之间的有效对抗将激励生成器合成更优质量的图像[55]。

2.2.1 生成对抗网络

GAN是一种通过博弈学习方法进行网络训练的模型，如图6所示，通过使用生成器(Generator)和判别器(Discriminator)进行相互博弈学习，共同提高网络各自的能力，使生成器能够模拟输入数据样本的分布，合成越来越真实的图像，直到判别器不能将生成器合成的样本和真实样本区分开来。

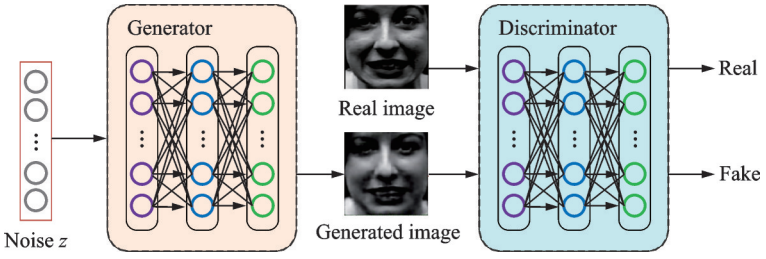


图6 生成对抗网络

Fig.6 Generative adversarial network

生成对抗网络中，生成器主要用于学习真实图像的分布，然后使用学习到的分布合成相似分布的数据，并让自身合成的数据更加真实，降低判别器的识别效果。而判别器的任务是判断当前图像是真实数据还是由生成器合成而来的数据。在整个流程中，生成器需要努力让合成图像更加真实，而判别器需要尽力去找出每一个合成的图像，这就是判别器和生成器的博弈学习。随着学习的深入，最终两个网络将达到一个纳什均衡：生成器能够合成十分接近真实数据分布的假数据，而判别器对于给定数据预测其真假的能力将达到最低，即正确预测当前数据的概率为0.5。据此，优化的目标函数定义为

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

式中：E代表数学期望；$D(x)$代表判别器D判别样本x属于真实分布的概率。

2.2.2 条件生成对抗网络

由于原始的GAN网络仅能够在无监督的条件使用随机噪声作为输入进行相关数据的合成，对于合成数据来说，其输出是随机、无法预测的。例如在人脸图像的相关合成中，无法对合成图像的相关参数进行控制，只能输出符合网络训练所使用的原始数据集分布的图像。因此，使用GAN进行有目的地合成图像或者数据成为当前研究的热点。条件生成对抗网络(Conditional GAN，CGAN)是一种条件式的生成对抗网络，通过给原始的生成对抗网络施加一个限制条件，能够合成有针对性、符合响应条件的图像或其他数据。

相对于原始的GAN，CGAN在网络的生成器中加入了约束条件y，使得生成器能够在约束条件y的限制下进行相应数据的合成，因此生成器合成的数据不再是随机的，而是变成了含有约束信息y的合成数据。CGAN的优化函数问题转化为

$$V(G,D)=E_{x\sim P_{\text{data}}}[\log D(x|y)]+E_{z\sim p_z(z)}[\log(1-D(G(z|y)))] \tag{2}$$

式中：$G(z|y)$代表基于约束条件y的合成图像；$D(x|y)$代表样本x属于真实分布的概率。

现有的表情合成研究工作中使用了不同的条件信息，如表情类别标签[56]、人脸几何信息[57]、人脸关键点[58]、AU标签和人脸轮廓图像[59]等，在进行图像合成时通过这些条件信息来指导图像的合成。基于条件生成对抗网络的GAN变体有很多，例如，Liu等[18]提出了一个基于3D人脸AU的条件对抗合成模型，将目标AU标签y_{target}作为条件辅助信息输入到合成模型，合成指定身份信息的目标表情参数$x_{\text{target}}^{\text{exp}}$。人脸几何信息可以客观地反映人物的表情信息及面部基础信息，因此一些学者将人脸几何信息输入CGAN中作为条件信息来指导表情的合成[25,60-63]。Huang等[27]提出了一个DyadGAN网

络,在面试官和面试者两个人的互动场景中,将面试者的表情轮廓特征图作为条件信息来辅助合成面试官的表情图像,试图模拟互动场景中一个人的表情对另一个人的表情响应的关系和影响。文献[26]提出的GC-GAN中,首先利用对比学习方法来学习目标表情的关键点特征图,从而削弱了不同脸型不对齐问题的影响;然后将关键点特征图作为条件信息,与输入图像的特征向量串联,经解码器合成带有目标表情的图像。文献[64]将目标表情标签作为条件信息,与输入图像一起作为输入,引导图像的合成。Ma等[62]提出了一种基于姿态的生成网络(Pose guided person generation network,PG^2),可以合成任意姿态下的图像。目标姿态由一组18个关节位置定义,编码为热图,将目标姿态的热图作为条件辅助信息,与PG^{2}中的输入图像串联输入到网络中。此外,又采用了两阶段合成方法来提高合成图像的质量。

2.2.3 生成对抗网络变体

本节总结了基于GAN的表情合成算法的基础网络结构,如图7所示,可分为3种情况。首先,网络的输入为条件信息和源图像,条件信息的种类如第1节提到的人脸特征信息,图7中条件信息示例分别为目标表情参考图、人脸轮廓图、人脸关键点和目标表情类别向量,研究时选择其一作为网络的条件信息即可。图7(a)将条件信息和源图像一起输入编-解码结构的生成器G中,输出生成图像;再将生成图像输入到判别器D中,判别图像的真假。图7(b)在生成器中设置两个编码器,分别提取条件信息中的目标表情特征和源图像的人物身份特征;然后将目标表情特征与身份特征级联,输入解码器中输出生成图像。图7(c)首先利用编码器提取源图像的身份特征,再与条件信息特征向量级联,解码输出生成图像。GAN算法的改进主要是对生成器或判别器的设计进行创新,代表性算法总结如下。

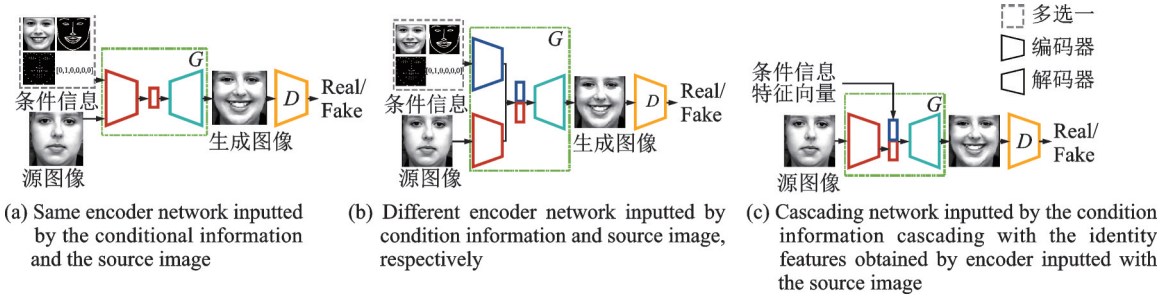


图7 基于GAN的表情合成算法基础网络结构

Fig.7 Basic networks of expression synthesis algorithm based on GAN

(1) StarGAN

Choi等[64]提出了一个星型网络框架StarGAN,与最初的GAN相比,StarGAN包括2个参数共享的生成器,形成CycleGAN[65]的环形式,并在网络中加入掩码向量,用来学习多个数据集之间的多域图像转换。该方法使StarGAN能够控制所有可用的域标签完成跨数据集的训练。而现有的基于GAN的模型若要实现在k个不同的域上进行属性的迁移,需要构建$k\times(k-1)$个生成器,并且还不能跨数据集训练。StarGAN算法可以用一个生成器来学习多个域之间的图像转换。该算法输入源为人脸图像和目标表情标签,经生成器合成图像;再将合成的图像串联源人脸图像的表情标签作为输入,经生成器重构出源图像。这两次合成过程成为一个环结构,计算源图像和重构的源图像之间的误差,作为损失函数的一部分来控制网络的合成效果。

(2) G2-GAN

Song等[25]提出了一种几何引导的生成对抗网络G2-GAN,该网络包括两个GAN结构,形成环的结

构。与StarGAN不同的是,StarGAN中的两个生成器是参数共享的同一个生成器,而在G2-GAN中,两个GAN结构是不同的,二者需要完成两个相反的任务:表情合成和表情移除。两个网络在中性表情和任意表情之间形成映射循环。G2-GAN将人脸关键点作为条件,指导合成具有特定表情、特定身份的人脸表情。训练G2-GAN时和StarGAN的环结构类似,输入的中性人脸经第1个GAN合成带有表情的人脸图像;再将合成的表情图像输入第2个GAN重构出中性人脸图像。计算源中性人脸图像与重构出的人脸图像的差异,作为重构损失来提升合成图像的效果。

(3) FaceID-GAN

Shen等[66]提出了一个三元GAN网络FaceID-GAN。原始的GAN是生成器与判别器之间的二元对抗,FaceID-GAN与原始的GAN不同,通过将人脸身份分类器P作为网络结构的第3个参与者,人脸身份分类器P和判别器D协同对抗生成器G,使得G合成质量较高,并保留人脸的身份信息的图像。FaceID-GAN不是简单地将身份分类器作为一个附加的鉴别器,而是通过满足输入数据信息对称性来构造FaceID-GAN,从而保证真实图像和合成图像投影到相同的特征空间,减小了网络的训练难度。

(4) Warp-Guided GAN

Geng等[67]提出了一种利用单幅照片合成实时人脸动画的Warp-Guided GAN框架。该框架包含2个具有不同任务的GAN:第1个GAN用来精细化人脸图像细节;第2个GAN用来合成口内牙齿区域。该方法只需要1幅人像图像和1组从驱动源(如照片或视频序列)派生出来的人脸关键点就可以合成具有丰富人脸细节的动画图像。该方法的核心是一个基于图像变形的合成模型,该模型分为3个阶段。给定1幅人像和1个驱动源(照片或视频),首先用追踪特征标记点对图像进行图像变形。然后裁剪出变形后的图像的人脸区域,计算出关键点的位移图,将变形后的人脸区域和关键点位移图输入精细化生成对抗网络WGGAN,合成逼真的面部细节,如皱纹、阴影等。精细化后的人脸输入另一个生成对抗网络HRH-GAN,合成适合目标表情的内口区域。最后将精细化的人脸无缝集成到扭曲的图像中,合成动画效果。

(5) Cascade expression focal GAN (Cascade EF-GAN)

Wu等[68]提出了一种AU标签引导的Cascade EF-GAN网络。该网络由4部分组成:1个全局生成器和3个局部生成器,分别为眼部生成器、鼻子生成器及嘴部生成器。将目标表情的AU向量分别与输入原图、眼部图像、鼻部图像和嘴部图像串联,构成4个子网络的输入。经过4个子生成器,分别生成带有目标表情的全人脸图像、眼部图像、鼻部图像和嘴部图像。然后将生成的3部分局部图像拼接成不完整的人脸图像,再与生成的全人脸图像级联,解码生成最终的目标表情人脸图像。该方法设计了一种级联的渐进生成策略,将一个变化大的表情分割为几个小的级联变换,这样在处理较大表情变换时,有助于抑制伪影,产生更真实的生成效果。

2.3 损失函数

损失函数是深度学习中最基础也最为关键的一部分。通过最小化损失函数,使模型达到收敛状态,减小模型预测值的误差。使用不同的损失函数,对模型的影响很大。在表情合成研究中,损失函数有对抗损失、像素损失、感知损失、循环一致性损失和三元组损失等,可根据模型情况选择或设计不同的损失函数来训练模型。

2.3.1 对抗损失

在基于GAN的表情合成模型中,对抗损失是必不可少的。GAN通过对抗损失使生成器和判别器相互对抗博弈,不断优化网络。原始对抗损失如式(1)所示。GAN的训练极其困难且不稳定,很难平衡

生成器与判别器的训练程度。式(1)使用Sigmoid交叉熵损失函数,容易造成梯度消失问题,使生成器的训练不充分。

基于此,Mao等[69]提出最小二乘GAN(Least squares GAN,LSGAN)中对GAN的损失函数进行改进,如式(3~4)所示,采用最小二乘损失函数代替Sigmoid交叉熵,缓解梯度消失问题。

$$\min_D V_{\mathrm{LSGAN}}(D)=\frac{1}{2}E_{x\sim p_{\mathrm{data}}(x)}[(D(x)-b)^2]+\frac{1}{2}E_{z\sim p_z(z)}[(D(G(z))-a)^2] \tag{3}$$

$$\min_G V_{\mathrm{LSGAN}}(G)=\frac{1}{2}E_{z\sim p_z(z)}[(D(G(z))-c)^2] \tag{4}$$

式中:常数a、b分别表示真实分布和生成分布的标注值;常数c表示在生成器优化时生成图像被判定为真实图像的标注值。

LSGAN在判别器的输出层中去掉sigmoid激活函数,且在损失函数中去掉log函数,使用最小二乘损失函数。这样使D不仅判别真假,同时惩罚离群的样本点,使生成样本不断向真实分布靠近。

Arjovsky等[70]为了解决GAN训练不稳定及生成器与判别器很难训练平衡的问题,提出WGAN(Wasserstein GAN)网络,同样对损失函数进行改进,使用Wassertein距离来衡量合成图像与真实图像分布之间的距离,具体公式如下

$$L_G=-E_{x\sim P_g}[f_w(x)] \tag{5}$$

$$L_D=E_{x\sim P_g}[f_w(x)]-E_{x\sim P_r}[f_w(x)] \tag{6}$$

式中:P_g代表生成分布;P_r代表真实分布;f_w代表满足Lipshitz连续条件的判别器。

LSGAN和WGAN改进后的损失函数提高了GAN的性能。此外,还有一些方法采用改进的DCGAN[71]、WGAN-GP[72]和BEGAN[73]损失函数使模型拥有更快的收敛速度,解决模式崩溃问题,或提升样本的生成质量。

2.3.2 像素损失

像素损失计算合成图像与目标图像的像素间损失。一般地,采用L-P范数$\|\cdot\|_p$计算两者差值,表达式为

$$L_{\mathrm{pix}}=\left\|t-G(x)\right\|_p \tag{7}$$

式中t代表目标图像。

2.3.3 感知损失

在计算2幅图像的像素损失时,若移动了某个像素点,或相同的图像使用不同的分辨率时,会得到一个很大的像素损失,这是不合理的。因此,研究人员提出使用预训练模型分别提取合成图像和真实图像的特征向量,计算二者的特征向量的差值得到感知损失。一些经典的深度神经网络的预训练模型被用来提取特征,如VGGNet、ResNet、InceptionNet和AlexNet等,具体公式为

$$L_{\mathrm{percept}}=\left\|\mathrm{VGG}(x_r)-\mathrm{VGG}(G(x))\right\| \tag{8}$$

式中:x_r代表真实图像;VGG(·)代表预训练的VGG网络,可替换为其他网络结构,公式也可设计为如图8所示的计算多层感知损失的形式。

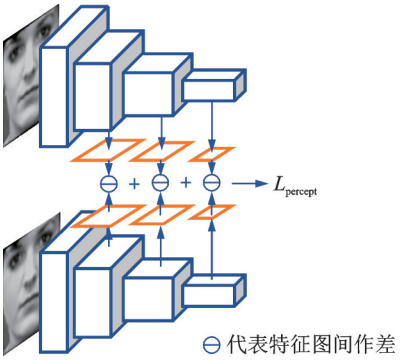


图8 多层感知损失计算方法

Fig.8 Calculation method of multi-layer perceived loss

2.3.4 循环一致性损失

受CycleGAN[65]的启发,研究人员将循环一致性损失应用于人脸合成领域。在表情转换的过程中,应保证人脸的身份信息不

发生变化，循环一致性损失可以强制网络在改变表情的同时保留人脸的身份信息。循环一致性损失的计算过程如图9所示，具体计算公式如下

$$L_{\text{cycle}} = \left\| x - G(G(x|c_1)|c_0) \right\|_{\text{p}} \tag{9}$$

式中：c_0代表源图像x对应的表情条件；c_1代表目标表情条件；$G(x|c)$代表基于表情条件c生成的合成图像。

2.3.5 身份保留损失

在表情合成过程中，需设置身份保留损失来保证身份信息的不丢失。如式(10)所示，一般采用预训练的人脸识别模型提取合成图像与输入图像的身份特征，并计算二者的差值。

$$L_{\text{id}} = \left\| \Theta(x) - \Theta(G(x)) \right\| \tag{10}$$

式中Θ为预训练的人脸识别模型。

2.3.6 三元组损失

为了提升合成图像的质量，三元组损失被提出来训练网络，使网络有更好的性能。文献[74]认为网络完成从输入域到输出域的转换后，不能使输出图像拥有和输入图像相同的分布。因此，设计了两种方式生成目标图像，直接生成和渐进生成，如图10所示。在渐进生成的方法中，首先将输入人脸正面化，将生成的正面化图像作为中间过渡图像，之后使用生成的过渡图像作为网络的输入，生成目标表情图像。然后计算两种方式的生成图像之间的差异，作为三元组损失。这个过程使生成图像作为网络的输入，保证了输入域与输出域的分布一致性。

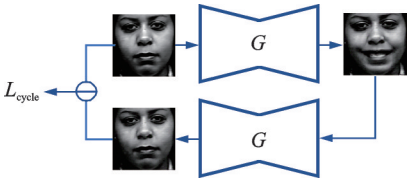
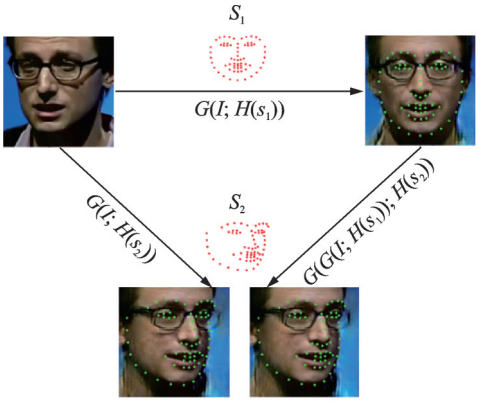


图9 循环一致性损失的计算

Fig.9 Calculation of the cycle consistency loss



$\mathcal{L}_{\text{triple}} = \|G(I; H(s_2)) - G(G(I; H(s_1)); H(s_2))\|$

图10 文献[74]提出的三元组损失

Fig.10 Triple loss proposed in Ref.[74]

2.4 代表性算法总结

表2总结了9种深度学习方法的性能和所用的数据库，并简单描述了算法特点和损失函数的设置情况。表2中“客观评价指标”字段中的下标“1”和“2”与“数据库”字段中的下标“1”和“2”相对应，表示在该数据库上的实验精度。算法应用时，根据自身算法的设计选择合适的损失函数或设计新的有效损失函数，同时根据本算法需要何种人脸标签、是否需要成对数据集、需要单幅图像或视频序列等选择合适的数据集进行训练。在表情合成领域，关于模型的评价标准尚不统一，因此很难判断哪个算法是最好的。除了对生成图像进行主观的视觉效果评价外，研究人员通常在人脸识别率、表情识别率和图像生成质量这3个方面对模型进行客观评估，评估方法在后文阐述。

表2 人脸表情合成领域代表性算法对比总结

Table 2 Comparison of the state-of-the-art facial expression synthesis algorithms

方法	发表时间	简单描述	损失函数	数据库	客观评价指标
EF-GAN[68]	CVPR 2020	AU引导;局部生成+全局生成+级联渐进生成+4个子网络	对抗损失+分类损失+身份保留损失+注意力损失+插值损失	$RaFD_1$ $CFEED_2$	表情分类精度:$93.67\%_1$, $89.25\%_2$ PSNR: 23.07_1, 21.34_2 FID: 42.36_1, 27.15_2
FReeNet[75]	CVPR 2020	关键点引导;关键点转换器+几何引导的生成器	对抗损失+像素损失+循环一致性损失+三元组感知损失	$RaFD_1$ $Multi\text{-}PIE_2$	FID: 12.17_1 SSIM: 0.717_1 AMT: $74.9\%_1$
FLNet[76]	AAAI 2020	关键点引导;语音人脸合成+双流网络+掩码融合	对抗损失+像素损失+感知损失+全变分损失	$TCD\text{-}TIMIT_1$ $FaceForensics_2$	L1 loss: 7.99_1, 10.20_2 FID: 17.07_1, 20.62_2
Gu[77]	CVPR 2019	目标表情掩模引导;属性编辑+掩模分块+3个子网络	对抗损失+局部像素损失+全局像素损失+掩模损失	Helen Dataset VGGFace2	FID: 8.92
Fan[78]	AAAI 2019	表情特征向量引导;用户可控表情强度+嘴部判别器+全局判别器	对抗损失+像素损失+关键点损失+相邻帧相似损失	CK+ CK++	SSIM: 0.953 AMT: >50%
StarGAN[64]	CVPR 2018	类别标签引导;单一网络多域图像翻译+星型网络+掩模向量多数据集训练+域分类	对抗损失+域分类损失+循环一致性损失	CelebA RaFD	AMT:各属性编辑得分高于对比方法 表情分类误差:2.12%
GANimation[79]	ECCV 2018	AU引导;注意力机制+注意力掩模生成器+颜色掩模生成器+掩模融合	对抗损失+注意力损失+AU分类损失+循环一致性损失	EmotioNet RaFD	
ELEGANT[80]	ECCV 2018	表情特征向量引导;交换目标属性生成	对抗损失+重建像素损失	CelebA	FID: 24.88 (best)
DyadGAN[27]	CVPR 2017	表情特征向量+轮廓图引导;两阶段生成+轮廓图驱动	对抗损失+像素损失	大学生入学面试视频数据库	假设检验;表情的类内距离和类间距离

3 人脸数据集

深度学习中数据集必不可少。在选择数据集前,需要了解数据集的数据分布情况,判断其是否与使用算法相匹配。本节总结了常用人脸数据集及其特征如下。

(1) CelebA数据集

CelebA是一个大型人脸属性数据集[81],包含10 177位名人的202 599幅图像,每幅图像带有人脸的5个关键点坐标信息和40个人脸属性的标签信息,例如是否带眼镜、有无刘海、是否戴帽子、是否卷发以及是否微笑等属性。数据集中的人脸图像有较大的姿势变化和杂乱的背景。CelebA具有多样性、数量多和注释丰富等特点,可以用作计算机视觉任务的训练集或测试集,如人脸属性识别、人脸检测、人脸关键点检测、人脸编辑和人脸合成等。

(2) Cohn-Kanade数据集

Cohn-Kanade数据集目前有CK和CK+两个版本,主要用于人脸图像分析与人脸合成的研究。

CK包含了97个志愿者的486个表情视频序列,这些志愿者被要求进行23个表情的展示,表情包括单个AU的运动或多个AU组合的面部肌肉的运动。每个视频序列都从中性表情开始,表情的表现力逐渐增大,一直达到表情程度的峰值结束。其中,中性表情是指人脸面无表情的状态。每个序列的峰值表情经FACS编码,得到各自的表情标签。

CK+[82]是CK基础上的扩展,是一个具有代表性的人脸表情数据集。CK+中志愿者量增加到123位,表情序列数增加到593个。与CK相同,593个视频序列中,每个序列的峰值表情都带有FACS编码的AU标签,其中327个序列同时带有表情标签,如快乐、悲伤、惊讶和愤怒等。

(3) Oulu-CASIA数据集

Oulu-CASIA数据集采用2个成像系统NIR (Near infrared)和VIS (Visible light),在3种不同光照条件下分别拍摄了80名志愿者的6种典型表情的视频。3种光照方式为正常的室内照明、弱照明(只有电脑显示屏的光)和暗照明(所有灯都关掉)。6种表情为快乐、悲伤、惊讶、愤怒、恐惧和厌恶。该数据库可用于研究光照变化对人脸表情、表情识别或人脸识别的影响。

(4) DISFA数据集

DISFA数据集[83]包含了27个不同种族的年轻人的脸部视频,这些视频是在志愿者观看视频片段时拍摄的,目的是捕捉志愿者自发的情绪表达。根据FACS,视频帧对应12种AU,分别为AU1、AU2、AU4、AU5、AU6、AU9、AU12、AU15、AU17、AU20、AU25和AU26。其中,每个AU的强度范围为0~5,每个视频帧中的AU强度值都是由专家手工标注。此外,此数据集还包括每幅图像的66个人脸关键点。

(5) RaFD数据集

RaFD数据集[84]是Radboud大学Nijmegen行为科学研究所整理的一个高质量的人脸数据集。此数据集中共包含49个模特,其中20名白人男性成年人,19名白人女性成年人,4个白人男孩,6个白人女孩。数据集中共有8 040幅图像,包含8种表情,即快乐、悲伤、惊讶、愤怒、恐惧、厌恶、蔑视和中立。每一个表情包含3个不同的注视方向,且使用5个相机从不同的角度同时拍摄。

(6) PIE数据集

PIE数据集[85]发布于2002年,包含68位志愿者的41 368幅图像,每个人有13种姿态,43种光照条件和4种表情。其中,姿态和光照变化也是在严格控制的条件下采集的,它在推动多姿势和多光照的人脸识别研究方面具有非常大的影响力,不过仍然存在模式单一多样性较差的问题。为了解决这些问题,卡内基梅隆大学的研究人员在2010年建立了Multi-PIE数据集[86]。此数据集包含337位志愿者,在15个视角、19个光照条件和不同的表情下记录,约75万幅图像。这些图像在头部姿势、光照和面部表情方面有很大的变化。此外,还获得了高分辨率的正面图像。

(7) RAF-DB数据集

RAF-DB数据集[87]是一个大型的人脸表情数据集,共包含29 672幅真实世界中的图像。此数据集中的图像在志愿者的年龄、性别、种族、头部姿势、光照条件、遮挡状况(如眼镜、面部毛发等)以及后处理操作(如滤镜等特殊效果)等方面有很大的变化。此外,该数据集有丰富的注释,图像中有7种基本表情和12种复合表情,并且每幅图带有5个精确的人脸关键点、年龄范围和性别标注。

表3为以上各数据集的对比表格,包括各数据集的样本量、采集情况、标签信息以及志愿者的表情是否摆拍等信息,研究人员可根据模型需求选择合适的数据集进行实验。有关人脸的数据集还有很多,例如:人脸检测方向的数据集有AFW、FDDB、WIDER Face和MALF等;人脸关键点检测方向的数

据集有XM2VTS、LFPW、Helen、IBUG、AFLW、300W、MTFL、MAFL和WFLW等;人脸识别方向的数据集有FERET、Yale、CAS-PEAL、LFW、Pubfig、MSRA-CFW、FaceScrub、UMDFaces、MegaFace、MS-Celeb-1M、VGG Face、IMDB-Face、YouTube Faces和IARPA Janus等;有关人脸年龄与性别的数据集有FGNet、CACD2000、Adience、IMDB-wiki和MORPH等;有关人脸姿态的数据集有BIWI、Bosphorus、HPD、BIWI kinect、FaceWarehouse、TMU、UPNA和300W-LP等。

表3 人脸数据集对比总结

Table 3 Comparison of facial expression datasets

数据集	发布年份	样本量	人数	简单描述	标签	自发/摆拍	环境
CelebA	2016	202 599I	10 177	大姿态变化+背景混杂	5个人脸关键点坐标+40种人脸属性二值标签(有/无)+身份标签		Wild
CK	2000	486V	97	23种表情	AU+表情类别标签	摆拍	实验室
CK+	2010	593V	123	23种表情	AU+表情类别标签+人脸关键点坐标	自发+摆拍	实验室
Oulu-CASIA		2 880V	80	6种表情视频+3种光照方式	光照+身份+情绪	摆拍	实验室
DISFA	2013	130 000I	27	每位志愿者采集4 845个视频帧	AU+人脸关键点坐标	自发	实验室
RaFD	2010	8 040I	49	8种表情+3个注视方向+5个不同角度拍摄	情绪类别+注视方向+头部姿势	摆拍	实验室
PIE	2000	41 368I	68	每位志愿者有13个角度+43种光照+4种表情	角度+光照+表情	摆拍	实验室
Multi-PIE	2009	755 370I	337	15个角度+19种光照+6种表情	角度+光照+表情	摆拍	实验室
RAF-DB		29 672I		单标签子集(7类情绪)+双标签子集(12类情绪)	情绪类别+5个人脸关键点坐标(精确标注)+37个人脸关键点坐标(自动标注)+种族+年龄+性别	自发+摆拍	Web

注:表中样本量一列中"I"表示图像,"V"表示视频。

4 实验评估方法

表情迁移的目的是生成带有指定表情的目标人脸逼真图像,有3个基本要求:保留人脸的身份信息、指定人脸做出目标表情及图像质量逼真自然,表情迁移的效果可以对这3点进行评估。身份信息的保留采用人脸识别率来评估;目标表情的表现力可以采用表情识别率和AU激活程度来评估;图像质量可以采用结构相似性指数来评估。

4.1 人脸识别率

人脸识别是基于人的脸部特征信息识别出该人脸的身份信息。将表情迁移生成的人脸图像进行身份信息的识别,可以有效地评估生成图像对身份信息的保留效果。人脸识别算法主要包括人脸检测、特征提取、人脸特征比较和输出识别结果4个步骤[88]。首先,人脸检测通过对采集到的包含人脸的图像进行人脸区域的获取,即在图像中准确描述人脸的区域和位置、大小等信息。然后,对人脸进行特征提取,再进行人脸的特征比较。人脸特征比较是对提取到的人脸特征进行特征距离的对比,即对获

取到的人脸特征进行分类,得到特征的分类结果。首先设定一个特征距离阈值,特征距离超过阈值将认为是同一个人的不同面部图像,据此获取到识别结果,并进行识别率的计算。在特征距离的计算中,通常使用欧式距离、马尔可夫距离以及卡方距离等,相关的分类器有SVM、神经网络(Nerual network, NN)以及HMM等[89]。

在人脸识别率的计算中,假设数据集中共有N个人的n幅人脸图像,其中,第i个人有M_i幅人脸图像,即

$$n=\sum_{i=1}^{N}M_i \tag{11}$$

则对于人脸识别系统Π,其识别率为

$$R=\frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{M_i}\Pi(\boldsymbol{I}(i,j))\times 100\% \tag{12}$$

式中$\boldsymbol{I}(i,j)$为第i个人的第j幅图像。

$$\Pi(\boldsymbol{I}(i,j))=\begin{cases}1 & \text{识别正确}\\ 0 & \text{识别错误}\end{cases} \tag{13}$$

4.2 表情识别率

若要判断生成图像的表情表现力,可利用一个表情识别模型来判断生成的图像,若图像能被正确分类到目标表情类别中,即可说明生成图像的表情表现力。

表情识别系统主要有4个基本部分组成:表情图像获取、表情图像预处理、表情特征提取以及表情分类识别。在StarGAN中[64],Choi等采用ResNet-18在RaFD数据集上训练了一个人脸表情分类器,达到了99.55%的准确率。然后使用训练好的分类器来评估实验结果,让分类器来判断生成图像的表情类别,并计算出最终的分类误差,作为实验的客观评价。误差越小,说明StarGAN生成的图像的表情越真实,表情表现力越好,越容易被表情分类器正确分类。

4.3 AU激活程度

表情AU是Parke定义的用于度量视频中人脸表情变化时脸部不同位置形变的一种方法。AU能够提供一整套描述人脸动作的参数,通过对人脸中不同区域响应的描述,能够使用简单的组合产生数量巨大的人脸表情。另外,由于AU定义了人脸中的大部分区域,可以作为人脸识别中用于归一化的方式。

AU在表情合成中的作用主要是对表情视频的真实性进行验证,可以使用合成视频同原始驱动表情进行表情AU区域的激活对比。对于原始的表情视频来说,计算合成视频的对应AU激活区域以及对应区域的激活值,就能够得到表情特征的保留程度,从而得到当前视频合成框架的表情保留效果。

大多数方法是通过对人脸进行分块,然后对不同的分块进行归一化并计算AU的激活程度,但是存在分块区域不准确而导致AU识别失败的问题。由于神经网络对图像的遮挡以及光照等具有较强的鲁棒性,通过深度学习技术能够提取到精确的AU,因此人脸AU的识别准确率获得了巨大提升。例如,深度区域和多标签学习的面部活动单元检测(Deep region and multi-label learning, DRML)[17]是一个基于深度学习的面部动作单元提取框架,它基于深度学习技术对基于区域的面部AU进一步划分,提取不同区域的深度特征,然后组合形成不同的特征矩阵,对其进行多标签分类,从而获得精确的面部AU位置和其对应的激活值大小。

文献[18]采用AU估计器可以将模型生成的表情参数准确归类,这些AU估计器是由真实图像中

提取出的表情参数训练而来。AU估计器为支持向量回归(Support vector regression,SVR)和有序支持向量回归(Ordinal support vector regression,OSVR)。估计器将表情参数作为输入特征,可以估计出具有6个强度级别的AU强度。

4.4 峰值信噪比和结构相似性

峰值信噪比(Peak signal to noise ratio,PSNR)是一种最普遍和使用最广泛的全参考图像客观评价指标,是基于误差敏感的图像质量评价。在评估人脸生成方法时,PSNR通过计算生成图像和真实值之间的像素误差来评估算法性能,数值越大表示失真越小,表达式为

$$\mathrm{PSNR}=10\log_{10}\left(\frac{(2^b-1)^2}{\mathrm{MSE}}\right) \tag{14}$$

$$\mathrm{MSE}=\frac{1}{H\times M}\sum_{i=1}^{H}\sum_{j=1}^{W}(X(i,j)-Y(i,j))^2 \tag{15}$$

式中:MSE(Mean square error)表示生成图像Y和真实图像X的均方误差;H、W分别为图像的高度和宽度;b为每像素的比特数,一般取8,即像素灰阶数为256。

结构相似性(Structure similarity,SSIM)也是一种全参考的图像质量评价指标,由3个部分组成,分别是亮度、图像照明度和图像的对比度。这3个部分的乘积组成了两幅图像的结构相似性指标数值[90]。在SSIM中使用图像的均值作为亮度的估计,使用图像的标准差作为对比度的估计,并且使用图像的协方差作为结构相似程度的估计。但是将SSIM用于评估生成的人脸图像时,如果生成的图像和真实图像的色调有差异,SSIM得到的相似度得分会低于真实水平。

为了度量GAN生成图像的质量,文献[25-26,91]均采用SSIM和PSNR两个评价指标作为定量评价标准,客观地评价模型的性能。

5 其他客观评价方法

第4节总结了一些表情合成领域的公知客观评价指标,但在实际研究工作中,研究人员为了更加全面、多角度地评价模型性能,提出了一些不同的客观评价方法。所以,目前在表情合成领域中没有统一的指标来评估算法的好坏。即使采用相同的指标来评价各自的模型,若计算方法或提取特征的深度模型不同,也会导致结果不可比较。本节对现有文献中的评价方法进行总结。

5.1 基于人脸识别模型的身份信息保留度

一些研究使用训练好的人脸识别模型分别提取生成图像和原图像的身份特征,计算二者的相似度,验证是否为同一身份,以此评估模型的身份信息保留能力。文献[92]采用ResNet-18作为人脸识别模型,在CelebA和LFW数据集上进行人脸验证,人脸识别率越高,说明生成人脸的身份信息保留的越好。在FaceID-GAN中,Shen等[66]在MS-Celeb-1M数据集上训练了一个人脸识别模型,来计算输入人脸图像的身份特征和输出人脸图像的身份特征之间的相似性得分。在G2-GAN中,Song等[25]利用VGG-FACE和Light CNN两个人脸识别模型验证提出方法的有效性。

5.2 IS指标

IS(Inception score)指标用来综合衡量GAN生成图像的质量和多样性两个指标。在GAN中,希望条件概率$P(y_{\mathrm{B}}|x)$可以被高度预测,其中x表示给定的图像,y_{B}表示图像中的主要内容,即标签。可以被高度预测即希望能容易判别出图像中的主要内容。一般用熵值来描述随机性:如果一个随机变量是高度可预测的,就有较低的熵;如果是高度不可预测的,则有较高的熵。使用Inception网络对生成的图像进行分类,然后预测$P(y_{\mathrm{B}}|x)$,用该概率来反应图像的质量,概率值越高,说明Inception网络越有把握将

图像正确分类,即图像的质量越高。

图像边缘概率计算方法为

$$\int_z p(y_B|x=G(z))\mathrm{d}z \tag{16}$$

式中:z为随机噪声;$G(z)$为随机噪声z通过生成器得到的生成图像。

如果生成图像的多样性很好,那么预测标签y_B的分布则有很高的熵,从而导致准确预测y_B的结果更难。为了综合两个指标,通过使用KL-divergence计算IS的值为

$$\mathrm{IS}(G)=\exp(E_{x\in p}D_{\mathrm{KL}}(p(y_B|x)\|p(y_B))) \tag{17}$$

式中D_{KL}为KL-divergence的计算公式。

IS在一定程度上可以反映生成图像的质量以及多样性。但由于IS的计算过程中只考虑了生成样本,没有考虑真实数据,无法反映真实数据与样本之间的距离,因此存在一些问题。比如,数值受样本选取的干扰较大、不适合在内部差异较大的数据集上使用、无法区分过拟合等。

5.3 特征之间的相似性计算

5.3.1 弗雷歇距离

Xiao等使用弗雷歇距离来验证生成图像的质量[80]。弗雷歇距离得分(Fréchet inception distance score,FID)由Heusel等[93]于2017年提出并使用,是评估生成图像质量的度量标准,专门用于评估生成对抗网络的性能。FID从原始图像的计算机视觉特征的统计方面来衡量两组图像的相似度,计算方法如式(18)所示。首先利用Inception网络来提取特征,然后使用高斯模型对特征空间进行建模,最后求解两个特征之间的距离。较低的FID意味着图像有较高的质量和多样性。

$$\mathrm{FID}=\|\mu_r-\mu_g\|+\mathrm{Tr}(\boldsymbol{\Sigma}_r+\boldsymbol{\Sigma}_g-2\sqrt{\boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_g}) \tag{18}$$

式中:Tr代表矩阵的迹;μ_r代表真实图像深度特征的均值;μ_g代表生成图像深度特征的均值;$\boldsymbol{\Sigma}_r$代表真实图像深度特征的协方差矩阵;$\boldsymbol{\Sigma}_g$代表生成图像深度特征的协方差矩阵。

相比较于IS,FID对噪声有更好的鲁棒性。但是,FID基于特征提取,即依赖于某些特征的出现或者不出现,无法描述这些特征的空间关系。例如,用GAN去生成人脸,如果嘴巴长在眼睛上面,FID可能也会认为它是一幅效果较好的人脸图像。

5.3.2 余弦相似度

文献[94]通过计算真实图像特征与生成图像特征的余弦相似度来衡量源特征的残留程度。首先,由训练好的ResNet-18模型提取出源图像和生成图像的特征;然后,计算出两个特征之间的平均余弦相似度。相似度越小,说明源特征去除得越彻底。余弦相似度表示式为

$$\cos\theta=\frac{\phi(x)\cdot\phi(G(x))}{\|\phi(x)\|\times\|\phi(G(x))\|} \tag{19}$$

式中:ϕ代表预训练的ResNet-18模型;$\phi(x)$代表样本x对应的深度特征。

5.3.3 余弦距离

在FaceID-GAN中,Shen等[66]在LFW数据集上验证了提出模型的身份信息保留能力。该方法利用模型中的身份信息提取器提取出真实图像和生成图像的身份信息特征,然后计算出真实图像的身份信息特征与生成图像的身份信息特征的余弦距离,以此作为人脸验证的度量标准。

$$\cos\theta=\frac{\Psi(x_r)\cdot\Psi(Y)}{\|\Psi(x_r)\|\times\|\Psi(Y)\|} \tag{20}$$

式中:Y代表生成图像;Ψ代表身份信息提取器,$\Psi(\cdot)$代表样本对应的身份特征。

5.3.4 欧氏距离

DyadGAN中,Huang等[27]验证了在两个人互动时,一个人的表情对另一个人表情的关系和影响。Huang等创建了2个集合:第1个集合是由真实图像的特征构成;第2个集合是由生成图像的特征构成。其中,真实图像的特征和生成图像的特征由Emotient Facet SDK模型提取得到。然后计算2个集合中特征之间的欧氏距离,以此来测量真实图像与生成图像之间的差异,并采用统计学中的显著性检验来判别两个集合是否有显著差异。

Huang等在第2个实验中创建了另外的2个集合,不同于第1次实验,集合A是根据被面试者的8种表情生成的面试官的图像,集合B是被面试者的8种表情的真实图像。然后,分别计算集合A中的生成图像特征与集合B中8种表情图像特征的欧式距离。结果表明,只有当A中生成图像的源根据图像(被面试者图像)与B中图像的表情同属一类时,两组特征之间的平均欧式距离最小,即验证了两人在互动时表情具有相似性。欧氏距离表达式为

$$ED(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{21}$$

式中:X、Y分别代表真实图像和生成图像的特征图;n代表特征图的元素个数;x_i、y_i分别代表真实图像特征图和生成图像特征图的第i个元素。

5.4 用户调查

为了定量评估,Choi等[64]使用Amazon Mechanical Turk(AMT)平台进行用户调查来评估合成图像的质量。给定4幅合成图像,分别是由4种不同的方法生成,志愿者被要求根据合成图像的真实感、属性迁移的效果和源图像人脸的身份信息保留效果这3点来选择4幅图像中最佳的生成图像。Zhou等[53]创建了一个在线调查,共征集了112位志愿者从3幅由3种不同方法生成的、随机排序的图像中选出最符合数据集真实图像的图像。

5.5 其他方法

文献[95]分别提取出生成图像和输入图像的身份特征向量,然后计算两个向量的平均内容距离(Average content distance, ACD)来评价身份信息的保留程度。在FaceID-GAN中,Shen等[66]在IJB-A数据集上提出了一种验证生成图像质量的方法。该方法首先用判别器D判断生成图像为真的概率,得到生成图像的重构误差,将重构误差的倒数定义为置信度,也可在某种程度上描述生成图像的质量。文献[96]使用MS(Mode score)和FID作为评价指标来定量地评估生成人脸与游戏风格参考人物之间的相似度。

6 存在问题及研究展望

虽然GAN的出现有力地推进了图像合成任务的研究,越来越多的GAN变体用于表情合成,但仍存在几个问题需要考虑:

(1) 如何设计基于GAN的网络框架。随着GAN变体的增多,研究人员对GAN的生成器与判别器进行了改进。例如,Bao等[97]在生成器中设计了2个编码器;Shen等[66]设计了一个分类器与判别器共同对抗生成器;Zhang等[75]在编码器与解码器之间放置了若干层残差块等。同时,研究人员也在探讨如何将GAN与注意力机制和跳跃连接等模块相融合。因此,如何对GAN内部结构进行改进以生成更佳效果的人脸图像是需进一步研究的方向之一。

(2) 是否需要成对的数据集训练模型。在进行表情合成时,若算法为有监督的训练,需要计算驱动人脸的中性图像与表情图像之间的差值作为损失函数来引导目标表情人脸的合成。因此,需要成对的

数据集来训练模型。而对于一些无监督算法,可采用判别器或分类器,对生成图像和目标图像进行分类预测,并计算二者的类别差值作为损失函数,则不需要成对数据集。

(3) 如何提升人脸表情合成的图像质量。根据目前最新的研究结果来看,人脸表情合成图像的质量还有上升空间。研究人员可在图像的清晰度和人脸的细节生成方面继续提升合成图像的质量。人脸细节对于合成图像真实性的表达至关重要,不应伴随有奇怪的牙齿、不规则的皮肤纹理及杂乱的头发等细节问题。如何合成精细的人脸细节、减少不合理的伪影以保留源人脸的个性特征,仍是需要研究的方向。

(4) 如何个性化地合成人脸表情。文献[98]提出的模型可以合成合理自然的表情,但是合成的表情是非个性化的,合成的表情和目标人脸的真实值之间有一定的差异。因此,个性化地合成人脸表情是未来的一个研究方向。

(5) 如何设计更轻量级的模型。深度学习方法是高度数据依赖型的算法,它的性能通常随着数据量的增加而不断增强,而大量的数据需要在昂贵的GPU上进行训练。因此,如何设计更轻量级的模型简便快捷地合成人脸表情,降低成本,也是需要进一步研究的方向。

参考文献:

[1] 蒋斌,甘勇,张焕龙,等.非正面人脸表情识别方法综述[J].计算机科学, 2019, 46(3): 53-62.
JIANG Bin, GAN Yong, ZHANG Huanlong, et al. Survey on non-frontal facial expression recognition methods[J]. Computer Science, 2019,46(3): 53-62.

[2] 刘剑, 金泽群.基于深度学习的人脸表情迁移方法[J].计算机科学, 2019,46(S1): 250-253.
LIU Jian, JIN Zequn. Facial expression transfer method based on deep learning[J]. Computer Science, 2019, 46(S1): 250-253.

[3] CHEN Y C, SHEN X, LIN Z, et al. Semantic component decomposition for face attribute manipulation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2019: 9859-9867.

[4] YANG F, BOURDEV L, SHECHTMAN E, et al. Facial expression editing in video using a temporally-smooth factorization [C]// Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2012: 861-868.

[5] YANG F, WANG J, SHECHTMAN E, et al. Expression flow for 3D-aware face component transfer[J]. ACM Transactions on Graphics (TOG), 2011, 30(4): 60.

[6] MOHAMMED U, PRINCE S J D, KAUTZ J. Visio-Lization: Generating novel facial images[J]. ACM Transactions on Graphics (TOG), 2009, 28(3): 57.

[7] KEMELMACHER-SHLIZERMAN I, SANKAR A, SHECHTMAN E, et al. Being john malkovich[C]//Proceedings of European Conference on Computer Vision. Berlin, Heidelberg: Springer, 2010: 341-353.

[8] 王晓慧,贾珈,蔡莲红.基于小波图像融合的表情细节合成[J].计算机研究与发展, 2013, 50(2): 387-393.
WANG Xiaohui, JIA Jia, CAI Lianhong. Expression detail synthesis based on wavelet-based image fusion[J].Journal of Computer Research and Development, 2013, 50(2): 387-393.

[9] LI K, XU F, WANG J, et al. A data-driven approach for facial expression synthesis in video[C]//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2012: 57-64.

[10] KRAMER R S S, JENKINS R, BURTON A M. InterFace: A software package for face image warping, averaging, and principal components analysis[J]. Behavior Research Methods, 2017, 49(6): 2002-2011.

[11] GARRIDO P, VALGAERTS L, REHMSEN O, et al. Automatic face reenactment[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2014: 4217-4224.

[12] THIES J, ZOLLHOFER M, STAMMINGER M, et al. Face2face: Real-time face capture and reenactment of RGB videos [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2016: 2387-2395.

[13] PARKE F I. Computer generated animation of aces[C]//Proceedings of the ACM Annual Conference.[S.l.]: ACM, 1972: 451-457.

[14] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of Advances in Neural Information Processing Systems. USA: MIT Press, 2014: 2672-2680.

[15] 范懿文,夏时洪.支持表情细节的语音驱动人脸动画[J].计算机辅助设计与图形学学报, 2013, 25(6): 890-899.

FAN Yiwen, XIA Shihong. Towards expressively speech-driven facial animation[J]. Journal of Computer-Aided Design & Computer Graphics, 2013, 25(6): 890-899.

[16] EKMAN P, FRIESEN W V. Facial action coding system (FACS): A technique for the measurement of facial action[M]. Palo Alto, CA: Consulting Psychologists Press, 1978.

[17] ZHAO K, CHU W S, ZHANG H. Deep region and multi-label learning for facial action unit detection[C]//Proceedings of IEEE Computer Vision and Pattern Recognition. USA: IEEE, 2016: 3391-3399.

[18] LIU Z, SONG G, CAI J, et al. Conditional adversarial synthesis of 3D facial action units[J]. Neurocomputing, 2019, 355: 200-208.

[19] 谢金衡,张炎生.基于深度残差金字塔网络的实时多人脸关键点定位算法[J].计算机应用, 2019, 39(12): 3659-3664.
XIE Jinheng, ZHANG Yansheng. An algorithm for real-time multi-face landmark localization based on deep residual and feature pyramid neural networks[J].Journal of Computer Applications, 2019, 39(12): 3659-3664.

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

[21] EDWARDS G J, COOTES T F, TAYLOR C J. Face recognition using active appearance models[C]//Proceedings of European Conference on Computer Vision. Berlin, Heidelberg: Springer, 1998: 581-595.

[22] COOTES T F, EDWARDS G J, TAYLOR C J. Active appearance models[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2001(6): 681-685.

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

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

[25] SONG L, LU Z, HE R, et al. Geometry guided adversarial facial expression synthesis[C]//Proceedings of 2018 ACM Multimedia Conference on Multimedia Conference.[S.l.]: ACM, 2018: 627-635.

[26] QIAO F, YAO N, JIAO Z, et al. Geometry-contrastive GAN for facial expression transfer[EB/OL]. (2018-10-22)[2019-12-20]. https://arxiv.org/abs/1802.01822v2.

[27] HUANG Y, KHAN S M. Dyadgan: Generating facial expressions in dyadic interactions[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). USA: IEEE, 2017: 2259-2266.

[28] JO Y, PARK J. SC-FEGAN: Face editing generative adversarial network with user's sketch and color[EB/OL]. (2019-02-18)[2019-12-20]. https://arxiv.org/abs/1902.06838.

[29] WU W, ZHANG Y, LI C, et al. ReenactGAN: Learning to reenact faces via boundary transfer[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany:[s.n.], 2018: 603-619.

[30] QIAN S, LIN K Y, WU W, et al. Make a face: Towards arbitrary high fidelity face manipulation[C]//Proceedings of the IEEE International Conference on Computer Vision. USA: IEEE, 2019: 10033-10042.

[31] HAN X, GAO C, YU Y. DeepSketch2Face: A deep learning based sketching system for 3D face and caricature modeling[J]. ACM Transactions on Graphics (TOG), 2017, 36(4): 126.

[32] GUO Y, TAO D, YU J, et al. Deep neural networks with relativity learning for facial expression recognition[C]//Proceedings of 2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). USA: IEEE, 2016: 1-6.

[33] KOELSTRA S, PANTIC M, PATRAS I. A dynamic texture-based approach to recognition of facial actions and their temporal models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(11): 1940-1954.

[34] COHEN I, SEBE N, GARG A, et al. Facial expression recognition from video sequences: Temporal and static modeling[J]. Computer Vision and Image Understanding, 2003, 91(1/2): 160-187.

[35] HSU C W, CHANG C C, LIN C J. A practical guide to support vector classification[M]. Taibei, China:[s.n.], 2003.

[36] CAI J, CHANG O, TANG X L, et al. Facial expression recognition method based on sparse batch normalization CNN[C]//Proceedings of 2018 37th Chinese Control Conference (CCC). USA: IEEE, 2018: 9608-9613.

[37] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554.

[38] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J].Science, 2006, 313(5786): 504-507.

[39] AN F, LIU Z. Facial expression recognition algorithm based on parameter adaptive initialization of CNN and LSTM[J]. The

Visual Computer, 2020, 36(3): 483-498.

[40] VALSTAR M, PANTIC M. Fully automatic facial action unit detection and temporal analysis[C]//Proceedings of 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06). USA: IEEE, 2006: 149-149.

[41] MAHOOR M H, ZHOU M, VEON K L, et al. Facial action unit recognition with sparse representation[C]//Proceedings of Face and Gesture 2011. USA: IEEE, 2011: 336-342.

[42] WHITEHILL J, OMLIN C W. Haar features for FACS AU recognition[C]//Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR06). USA: IEEE, 2006: 97-101.

[43] ZHAO K, CHU W S, DE LA TORRE F, et al. Joint patch and multi-label learning for facial action unit detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2015: 2207-2216.

[44] LI W, ABTAHI F, ZHU Z, et al. Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection[C]//Proceedings of 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). USA: IEEE, 2017: 103-110.

[45] SHAO Z, LIU Z, CAI J, et al. Deep adaptive attention for joint facial action unit detection and face alignment[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany:[s.n.], 2018: 705-720.

[46] CORNEANU C, MADADI M, ESCALERA S. Deep structure inference network for facial action unit recognition[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany:[s.n.], 2018: 298-313.

[47] LI Y, ZENG J, SHAN S, et al. Self-supervised representation learning from videos for facial action unit detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2019: 10924-10933.

[48] MA C, CHEN L, YONG J. AU R-CNN: Encoding expert prior knowledge into R-CNN for action unit detection[J]. Neurocomputing, 2019, 355: 35-47.

[49] ZHANG Z, LUO P, LOY C C, et al. Facial landmark detection by deep multi-task learning[C]//Proceedings of European Conference on Computer Vision. Cham, Switzerland: Springer, 2014: 94-108.

[50] ZHANG K, ZHANG Z, LI Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.

[51] WU Y, HASSNER T, KIM K G, et al. Facial landmark detection with tweaked convolutional neural networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(12): 3067-3074.

[52] KOWALSKI M, NARUNIEC J, TRZCINSKI T. Deep alignment network: A convolutional neural network for robust face alignment[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. USA: IEEE, 2017: 88-97.

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

[54] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]//Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham, Switzerland: Springer, 2015: 234-241.

[55] 陈松, 袁训明. 动态人脸表情合成的模型特征驱动算法综述[J]. 计算机与现代化, 2019(7): 47-54.
CHEN Song, YUAN Xunming. A survey of dynamic human facial expression synthesis approach driven by model features[J]. Computer and Modernization, 2019(7): 47-54.

[56] MIRZA M, OSINDERO S. Conditional generative adversarial nets[EB/OL]. (2014-11-06)[2019-12-20]. https://arxiv.org/abs/1411.1784.

[57] REED S, AKATA Z, YAN X, et al. Generative adversarial text to image synthesis[EB/OL]. (2016-02-05)[2019-12-20]. https://arxiv.org/abs/1605.05396.

[58] FENG Y, WU F, SHAO X, et al. Joint 3D face reconstruction and dense alignment with position map regression network [C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany:[s.n.], 2018: 534-551.

[59] ISOLA P, ZHU J Y, ZHOU T, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2017: 1125-1134.

[60] BALAKRISHNAN G, ZHAO A, DALCA A V, et al. Synthesizing images of humans in unseen poses[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2018: 8340-8348.

[61] KOSSAIFI J, TRAN L, PANAGAKIS Y, et al. Gagan: Geometry-aware generative adversarial networks[C]//Proceedings

of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2018: 878-887.

[62] MA L, JIA X, SUN Q, et al. Pose guided person image generation[C]//Proceedings of Advances in Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc,2017: 406-416.

[63] WANG W, ALAMEDA-PINEDA X, XU D, et al. Every smile is unique: Landmark-guided diverse smile generation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2018: 7083-7092.

[64] CHOI Y, CHOI M, KIM M, et al. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2018: 8789-8797.

[65] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE International Conference on Computer Vision. USA: IEEE, 2017: 2223-2232.

[66] SHEN Y, LUO P, YAN J, et al. FaceID-GAN: Learning a symmetry three-player GAN for identity-preserving face synthesis [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2018: 821-830.

[67] GENG J, SHAO T, ZHENG Y, et al. Warp-guided GANs for single-photo facial animation[C]//Proceedings of SIGGRAPH Asia 2018 Technical Papers.[S.l.]: ACM, 2018: 231.

[68] WU R, ZHANG G, LU S, et al. Cascade EF-GAN: Progressive facial expression editing with local focuses[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2020: 5021-5030.

[69] MAO X, LI Q, XIE H, et al. Least squares generative adversarial networks[C]//Proceedings of the IEEE international Conference on Computer Vision. USA: IEEE, 2017: 2794-2802.

[70] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein GAN[EB/OL]. (2017-12-06)[2019-12-20]. https://arxiv.org/abs/1701.07875.

[71] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[EB/OL]. (2016-02-07)[2019-12-20]. https://arxiv.org/pdf/1511.06434.

[72] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of wasserstein gans[C]//Proceedings of Advances in Neural Information Processing Systems.[S.l.]: ACM, 2017: 5767-5777.

[73] BERTHELOT D, SCHUMM T, METZ L. BEGAN: Boundary equilibrium generative adversarial networks[EB/OL]. (2017-05-31)[2019-12-20]. https://arxiv.org/abs/1703.10717.

[74] SANCHEZ E, VALSTAR M. Triple consistency loss for pairing distributions in GAN-based face synthesis[EB/OL]. (2018-11-08)[2019-12-20].https://arxiv.org/abs/1811.03492.

[75] ZHANG J, ZENG X, WANG M, et al. FReeNet: Multi-identity face reenactment[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2020: 5326-5335.

[76] GU K, ZHOU Y, HUANG T S. FLNet: Landmark driven fetching and learning network for faithful talking facial animation synthesis[C]//Proceedings of AAAI. USA: AAAI, 2020: 10861-10868.

[77] GU S, BAO J, YANG H, et al. Mask-guided portrait editing with conditional gans[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2019: 3436-3445.

[78] FAN L, HUANG W, GAN C, et al. Controllable image-to-video translation: A case study on facial expression generation [C]//Proceedings of the AAAI Conference on Artificial Intelligence. USA: AAAI, 2019, 33: 3510-3517.

[79] PUMAROLA A, AGUDO A, MARTINEZ A M, et al. Ganimation: Anatomically-aware facial animation from a single image [C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany:[s.n.], 2018: 818-833.

[80] XIAO T, HONG J, MA J. Elegant: Exchanging latent encodings with GAN for transferring multiple face attributes[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany:[s.n.], 2018: 168-184.

[81] LIU Z, LUO P, WANG X, et al. Deep learning face attributes in the wild[C]//Proceedings of the IEEE International Conference on Computer Vision. USA: IEEE, 2015: 3730-3738.

[82] LUCEY P, COHN J F, KANADE T, et al. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression[C]//Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. USA: IEEE, 2010: 94-101.

[83] MAVADATI S M, MAHOOR M H, BARTLETT K, et al. Disfa: A spontaneous facial action intensity database[J]. IEEE Transactions on Affective Computing, 2013, 4(2): 151-160.

[84] LANGNER O, DOTSCH R, BIJLSTRA G, et al. Presentation and validation of the Radboud Faces Database[J]. Cognition and Emotion, 2010, 24(8): 1377-1388.

[85] SIM T, BAKER S, BSAT M. The CMU pose, illumination, and expression (PIE) database[C]//Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition. USA: IEEE, 2002: 53-58.

[86] GROSS R, MATTHEWS I, COHN J, et al. Multi-pie[J]. Image and Vision Computing, 2010, 28(5): 807-813.

[87] LI S, DENG W, DU J P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2017: 2852-2861.

[88] 胡步发, 黄银成, 陈炳兴. 基于层次分析法语义知识的人脸表情识别新方法[J]. 中国图像图形学报, 2018, 16(3): 420-426.
HU Bufa, HUANG Yincheng, CHEN Bingxing. A novel facial expression recognition method based on semantic knowledge of analytical hierarchy process[J]. Journal of Image and Graphics, 2018, 16(3): 420-426.

[89] 徐峰, 张军平. 人脸微表情识别综述[J]. 自动化学报, 2017, 43(3): 333-348.
XU Feng, ZHANG Junping. Facial microexpression recognition: A survey[J].Acta Automatica Sinica, 2017, 43(3): 333-348.

[90] BLANZ V, BASSO C, POGGIO T, et al. Reanimating faces in images and video[C]//Proceedings of Computer Graphics Forum. Oxford, UK: Blackwell Publishing, Inc, 2003, 22(3): 641-650.

[91] YUAN X, PARK I K. Face de-occlusion using 3D morphable model and generative adversarial network[EB/OL]. (2019-03-08) [2019-12-20]. https://arxiv.org/abs/1904.06109.

[92] ZHANG G, KAN M, SHAN S, et al. Generative adversarial network with spatial attention for face attribute editing[C]// Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: [s.n.], 2018: 417-432.

[93] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[C]//Proceedings of Advances in Neural Information Processing Systems.[S.l.]:[s.n.], 2017: 6626-6637.

[94] ZHU D, LIU S, JIANG W, et al. UGAN: Untraceable GAN for multi-domain face translation[EB/OL]. (2019-01-28)[2019-12-20]. https://arxiv.org/abs/1907.11418.

[95] GENG Z, CAO C, TULYAKOV S. 3D guided fine-grained face manipulation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2019: 9821-9830.

[96] SHI T, YUAN Y, FAN C, et al. Face-to-parameter translation for game character auto-creation[C]//Proceedings of the IEEE International Conference on Computer Vision. USA: IEEE, 2019: 161-170.

[97] BAO J, CHEN D, WEN F, et al. Towards open-set identity preserving face synthesis[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2018.

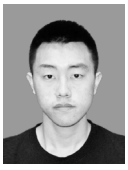
[98] WANG X, LI W, MU G, et al. Facial expression synthesis by U-net conditional generative adversarial networks[C]// Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval.[S.l.]: ACM, 2018: 283-290.

作者简介：

郭迎春(1970-),女,副教授,研究方向:图像处理与模式识别、人工智能,E-mail: gyc@scse.hebut.edu.cn。

王静洁(1995-),女,硕士研究生,研究方向:图像处理与模式识别。

刘依(1977-),**通信作者**,女,讲师,研究方向:图像处理与模式识别,E-mail: liuyi@scse.hebut.edu.cn。



夏伟毅(1997-),男,硕士研究生,研究方向:图像处理与模式识别。

张吉俊(1996-),男,硕士研究生,研究方向:图像处理与模式识别。

李学博(1996-),男,硕士研究生,研究方向:计算机视觉、图像处理与模式识别。

王天瑞(1996-),男,硕士研究生,研究方向:图像处理与模式识别。

(编辑:张黄群)