第 12 卷 第 4 期
2023 年 7 月
集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY
Vol. 12 No. 4
Jul. 2023

引文格式：
白泽琛, 姚乃明, 刘璐, 等. 基于加权混合融合变形的虚拟人情感表达 [J]. 集成技术, 2023, 12(4): 42-53.
Bai ZC, Yao NM, Liu L, et al. Blendshape-based emotional expressions generation for virtual human [J]. Journal of Integration Technology, 2023, 12(4): 42-53.

基于加权混合融合变形的虚拟人情感表达

白泽琛[1,2] 姚乃明[1] 刘 璐[1] 陈 鹏[1] 陈 辉[1*]

[1]（中国科学院软件研究所 人机交互北京市重点实验室 北京 100190）
[2]（中国科学院大学计算机学院 北京 100089）

摘 要 随着相关技术的发展，基于虚拟现实的人机交互技术越来越受到人们的关注。虚拟人作为一种直观的交互对象，在虚拟现实环境中扮演着十分重要的角色。在构建富有亲和力的虚拟人的过程中，制作虚拟人情感表达不可或缺。目前，主流的三维虚拟人情感表达主要依赖设计师手动制作面部表情动画，过程冗长，耗时费力。针对上述问题，该研究提出一种基于加权混合融合变形的虚拟人情感表达生成方法。该方法可以基于任意给定人脸表情图像，估计出三维虚拟人的目标混合形状（blendshape）的系数，进而自动化生成三维人类表情动画。实验结果表明，该方法具有较强的通用性和可迁移性，可有效减轻设计师制作情感表达面部动画时的工作量。

关键词 虚拟人；情感表达；虚拟现实；人机交互
中图分类号 TP 37 **文献标志码** A doi: 10.12146/j.issn.2095-3135.20221125001

Blendshape-Based Emotional Expressions Generation for Virtual Human

BAI Zechen[1,2] YAO Naiming[1] LIU Lu[1] CHEN Peng[1] CHEN Hui[1*]

[1](*Beijing Key Laboratory of Human-Computer Interaction*, *Institute of Software*, *Chinese Academy of Sciences*, *Beijing* 100190, *China*)
[2](*School of Computer Science and Technology*, *University of Chinese Academy of Sciences*, *Beijing* 100089, *China*)
[*]*Corresponding Author*: *chenhui@iscas.ac.cn*

Abstract With the rapid development of related technology, the virtual reality is attracting increasing attention. As an intuitive object for human-computer interaction, the virtual human plays an important role in the virtual environment. During the process of building plausible virtual human, one crucial step is to create emotional facial expressions. The mainstream methods often rely on hand-crafted efforts by designers, resulting a laborious and time-consuming task. To address this problem, this paper introduces a method that generates emotional expressions based on manipulating blendshapes. Given an arbitrary facial expression image, this method estimates the corresponding blendshape coefficients, which can be used to generate the target emotional expression on the

收稿日期：2022-11-25 **修回日期**：2023-02-28
基金项目：国家重点研发计划项目（2020YFC2004100）
作者简介：白泽琛，硕士，研究方向为虚拟现实和人机交互；姚乃明，博士，研究方向为人机交互；刘璐，硕士，研究方向为三维建模；陈鹏，硕士研究生，研究方向为人机交互；陈辉（通讯作者），研究员，博士生导师，研究方向为情感交互、人机交互和虚拟现实，E-mail：chenhui@iscas.ac.cn。

virtual human face. Experiment results show that the proposed method has strong generalization ability and is effective on reducing the burden of human designers in the task of emotional expressions generation.

Keywords　virtual human; emotional expression; virtual reality; human-computer interaction

Funding　This work is supported by National Key Research and Development Program of China (2020YFC2004100)

1　引　言

近年来，随着虚拟现实和人机交互技术的发展与成熟，人们越来越渴望更加自然的人机交互方式。由计算机创建的虚拟环境，具有逼真度高、临场感强等诸多优点，用户可沉浸在虚拟环境中，获得更为新奇、富有沉浸感的交互体验[1]。在虚拟环境中，虚拟人及其情感化的表达扮演着至关重要的角色[2]。由于虚拟人和人类的形象类似，天生富有亲和力，与人-物体交互相比，人类更倾向于和虚拟人进行交互，以获得更为丰富的交互体验[3]。和人类相似，虚拟人的情感表达有多种形式，包括面部表情、肢体动作、言语和音调等。其中，面部表情是最常见且最直观传递情感的形式之一，被应用于众多的虚拟现实系统[4]。因此，在构建虚拟人的过程中，为虚拟人制作生动形象的面部表情动画不可或缺，但手动制作虚拟人面部表情动画耗时耗力。目前，大部分虚拟现实应用中的虚拟人通过三维建模软件进行制作，如 Maya。在三维建模软件中，虚拟人的表情可通过加权混合融合变形得到，这意味着设计师需要通过试错的方式，不断地调整多个混合形状(blendshape)的权重系数，耗费大量时间，尤其制作大批量的表情动画时，这一弊端尤为突出。

为解决上述问题，本文提出一种方法，可以自动化估计人脸表情所对应的 blendshape 系数，方便制作人脸表情动画，减少生成虚拟人情感表达的工作量。具体地，本文选取人脸表情图像作为数据源，从图像中估算出一组 blendshape 系数，使得三维虚拟人模型能基于该组 blendshape 系数，复现二维图像上的表情。本文选取二维人脸图像作为数据源的优点有：人脸图像资源丰富，易于获取，且兼具较强的情感表达能力，因此，是非常理想的数据源。该方案的难点主要有：(1)源数据人脸表情图像中的身份、表情、姿态等属性是耦合在一起的，实现本文目标需要对人脸表情图像中的诸多属性进行解耦，并单独提取出表情特征，然后基于此进行 blendshape 系数估计；(2)该方法的预测估计目标，即 blendshape 系数，与虚拟人模型拓扑结构密切相关，而虚拟人的多样性为本方法带来了通用性层面的挑战。

为应对上述挑战，本文提出一种两阶段式的方案，对模型进行模块化设计。第 1 阶段：本文通过训练三维可变形模型(3D morphable model，3DMM)参数回归模型，获取二维图像中的人脸在三维空间中的参数化表示；第 2 阶段：本文提出一种 blendshape 系数估计模型，基于 3DMM 表情参数，估计目标虚拟人模型的 blendshape 权重系数。针对第 1 个难点，第 1 阶段的 3DMM 参数回归模型显式地将人脸参数解耦为身份、表情、姿态以及相机参数，使后续阶段的模型可以便捷有效地提取表情参数。针对第 2 个难点，本文模型对于不同虚拟人具有较强的可迁移性。这一特性主要体现在：第 1 阶段的 3DMM 参数回

归模型仅需一次训练，即可适用于任意三维虚拟人模型。对于第 2 阶段的 blendshape 系数估计模型，针对外观纹理不同，但 blendshape 拓扑结构相同的虚拟人，该模型仅需一次训练即可实现通用。针对 blendshape 拓扑结构不同的虚拟人，仅需对 blendshape 系数估计模型进行自动化微调更新即可完成迁移适用。

本文主要贡献如下：(1)针对虚拟人情感表达问题，本文提出了一套自动化的 blendshape 系数估计算法，实验与评估结果表明，该算法可有效减少设计师制作情感表达面部动画时的工作量。(2)在技术层面，该方法通过训练 3DMM 参数回归模型，有效缓解了人脸图像中属性耦合的问题。基于解耦后的人类表情参数，可有效支持虚拟人情感表达生成。(3)在应用层面，该方法具有较强的通用性和可迁移性，对于共享相同 blendshape 拓扑结构的虚拟人模型仅需一次训练即可通用；针对 blendshape 拓扑结构不同的虚拟人，对预训练模型的部分模块进行自动化微调更新即可完成迁移。

2 相关工作

虚拟人情感表达一般分为两大类：言语情感表达和非言语情感表达。其中，言语情感表达较为直接，一般基于文字或者说话[5]；非言语情感表达则较为复杂，涵盖内容也更为丰富，包括但不限于面部表情、眼神目光(又包括目光方向和瞳孔状态)、肢体动作、身体移动、身体距离、肢体接触和声音音调等[6-8]。研究表明，非言语表达能够引起用户的兴趣，触发某些动作，对于用户的行为有重要影响。除上述常见的非言语情感表达方式外，有研究者另辟蹊径，充分发挥虚拟环境的特性和优势，利用光照、阴影、线条和滤镜等表达情感状态[9]。本文的研究重点是虚拟人非言语情感表达中关于面部表情的情感化表达。

从技术上看，虚拟人非言语表达中关于面部表情的情感化表达实现方式一般有两种：二维图像驱动三维模型实现情感表达和纯三维方式实现情感表达。

2.1 二维图像驱动三维模型实现情感表达

在二维图像驱动三维模型实现情感表达的相关研究中，最相关的技术问题是三维人脸重建。三维人脸重建是计算机视觉中的经典问题，旨在基于二维人脸图像恢复三维的人脸信息。在三维人脸重建研究中，最具有代表性的工作之一是 Blanz 等[10]提出的 3DMM，该工作使用主成分分析(principal component analysis，PCA)对人脸的三维网格进行参数化，然后优化 PCA 的参数以拟合输入的二维人脸图像。在该工作基础上出现了一系列新的研究工作，旨在提升可形变模型在形状、纹理和表情上的表征能力。近年来，一系列工作将深度卷积神经网络引入三维人脸重建，借助高层级的图像语义表示，可预测可形变模型的参数。此外，生成对抗网络[11]也被应用于该领域，用于生成高逼真度的纹理。

基于 3DMM 的三维人脸重建，本质上可认为是一个二维人脸图像和三维人脸模型之间的参数拟合问题，可微分渲染器是解决这一问题的关键。近年来，随着相关技术的发展，可微分渲染器与自监督损失函数进行搭配，可有效训练神经网络进行参数拟合，无须扫描三维人脸数据[12]。然而，此类方法依旧无法直接解决本文中的 blendshape 系数估计问题。一方面，为不同 blendshape 拓扑结构的虚拟人分别训练可微分渲染器的开销巨大，不具备可推广性；另一方面，如何设计自监督损失函数也是一大难点。而 3DMM 的通用性与可靠性在学术界已被广泛验证，本文借用 3DMM 参数作为二维人脸在三维空间中的参数化表示，同时帮助显式地提取出表情参数。

另一个与二维图像驱动三维模型研究密切相关的领域是基于二维图像的游戏角色自动生成。

在计算机游戏和虚拟现实中，游戏角色自动生成扮演着至关重要的角色。近年来，游戏角色自动生成作为一种新兴技术，受到越来越多研究者的关注。2017 年，Wolf 等[13]首次提出一种对抗训练的方法——捆绑输出合成。该方法基于人脸图像创建参数化的虚拟化身，可利用对抗训练的方式，从预定义的人脸属性模板库中选取人脸属性。但人脸属性的选取与否是一个离散的数据，而非连续的人脸参数。2019 年，Shi 等[14]提出一种人脸生成参数方法可以基于人脸图像估计出一组连续的捏脸参数数值。该方法在第 1 阶段，训练一个生成模型，以实现可微分虚拟角色渲染；在第 2 阶段，借助额外的判别模型，度量和最大化输入图像和渲染虚拟角色图像的相似度，训练一个捏脸参数估计模型。但该方法通过一种迭代式搜索的方式进行捏脸参数估计，流程较为烦琐。2020 年，Shi 等[15]提出升级版的人脸生成参数，该方法通过自监督学习的方式，将捏脸参数估计的任务完全集成到一个“翻译模型”上，经过训练后，该“翻译模型”仅需一次神经网络的前向运算，即可完成捏脸参数估计，大大提升了效率。

本文的研究与游戏角色自动生成密切相关。相同点在于，二者均基于二维人脸图像，估计出一组人脸参数（离散的或连续的）。不同点在于，游戏角色自动生成需要估计出人脸几乎全部的参数，囊括了身份、表情和姿态，甚至纹理、发型等参数，使得创建的游戏角色最大限度地接近输入人脸图像；本文研究则重点关注对输入人脸图像的表情信息进行提取和变换，得到目标 blendshape 的权重系数，使得虚拟人能够合成接近输入人脸图像的情感表达。在技术方面，本文研究的问题需要对人脸图像中的表情信息进解耦和提取，这是本文研究独有的难点；在应用方面，游戏角色自动生成要求游戏角色含有更丰富的捏脸参数空间，以支持对身份、纹理等方面的改变，本文研究仅需虚拟人含有支持基本表情变化的 blendshape，通用度更高。

2.2　直接操作三维模型实现情感表达

有关使用直接操作三维模型生成基于面部表情的情感表达的研究，可以分为基于数据的情感表达和基于规则的情感表达两大类。

有研究人员选择基于表达行为数据库，从中自动地识别和提取情感化表达行为。研究人员通常使用动作捕捉的方法构建数据库。Emilya 数据库[16]采集了某个演员在各种情感状态下，执行一些简单任务时的全身动作以及面部表情的捕捉数据。笑声交互的多模态多人语料库[17]是采用人们在互动中大笑的动作捕捉数据构建的。基于特定的动作捕捉数据集，可利用机器学习技术识别和提取情感化表达的特征，并建立情感化表达的计算模型。2014 年，Ding 等[18]将机器学习技术应用于人类笑声的数据库，该计算模型学习了身体运动、面部表情和笑声声学特征之间的关系。基于数据的方法往往需要较大的数据量，数据收集的过程可能既困难又昂贵，但这类方法较为通用，能够不断加入新的数据。人文和社会科学的研究提出了关于人类如何表达情感的不同理论。1978 年，Ekman 等[19]提出了一个描述面部表情如何工作的模型——面部动作编码系统（facial action coding system，FACS），用于描述肌肉运动级别的面部表情。研究者通常使用 FACS 对情感化的面部表情进行编码。

随着现代数字产业的发展，基于 blendshape 构建情感化表达的方式也逐渐流行起来。blendshape 指对三维模型的单个网格变形，以实现许多预定义形状和任意数量形状组合的技术，常见于三维建模软件中，如 Maya、3D-Max 等。若制作三维人脸表情动画，设计师可以在三维建模软件中，根据一定的规则，甚至是个人的喜好，在三维人脸模型上制作一定数量的 blendshape，然后对多个 blendshape 进行混合，

可以使三维人脸模型表达出生动的表情[20-22]。基于三维建模软件的制作方式一般不需要数据驱动，生产成本较低。此外，可以通过定制规则，以满足特定需求（如场景、文化或性别特定行为），从而获得丰富的多模式行为库。然而，这类方法的弊端在于，组合 blendshape 的系数往往需要设计师通过手动试错的方式进行调整。

本文基于纯三维方式的 blendshape 构建情感表达，并借鉴了三维人脸重建和游戏角色生成领域中二维驱动三维的方式，其优势在于可以直接使用资源丰富的二维图像进行驱动生成。本文方法结合了二维驱动三维和纯三维方式的优势，既有二维驱动三维的资源易得性，又有直接操作三维实现情感表达的灵活性。因此，本文研究结合了基于数据的方法和基于 blendshape 的方法两者的优点，通过数据驱动的方式，提升了基于 blendshape 的虚拟人情感表达制作效率。

3 方 法

3.1 模型

本节提出一种基于二维人脸图像的 blendshape 系数估计算法。整个任务可以形式化地描述为：以真实的二维人脸图像作为目标表情，本算法估计一组 blendshape 系数，使其在预先定义的三维虚拟人模型上，可以重现目标面部表情。所提出的算法主要包含两个模块：3DMM 参数回归模块，负责从给定人脸图像中回归一组 3DMM 参数；blendshape 系数估计模块，将 3DMM 参数转换为适合目标三维虚拟人拓扑结构的 blendshape 系数。

3.1.1 3DMM 参数回归模型

如图 1 所示，本节首先采用一个基于 ResNet-50[23]的神经网络模型，以回归 3DMM 的参数（身份、表情、纹理）和相机参数（姿态、光照）。在 3DMM 参数体系下，人脸的形状 $\boldsymbol{S}$ 和纹理 $\boldsymbol{T}$ 如公式（1）～（2）所示。

$$\boldsymbol{S}=S(\beta,\gamma)=\overline{\boldsymbol{S}}+\boldsymbol{B}_{\mathrm{id}}\beta+\boldsymbol{B}_{\mathrm{exp}}\gamma \tag{1}$$

$$\boldsymbol{T}=T(\delta)=\overline{\boldsymbol{T}}+\boldsymbol{B}_{\mathrm{t}}\delta \tag{2}$$

其中，$\overline{\boldsymbol{S}}$ 为平均人脸的形状；$\overline{\boldsymbol{T}}$ 为平均人脸的纹理；$\boldsymbol{B}_{\mathrm{id}}$ 为身份的 PCA 基；β 为身份 PCA 基对应的参数（简称身份参数）；$\boldsymbol{B}_{\mathrm{exp}}$ 为表情的 PCA 基；γ 为表情 PCA 基对应的参数（简称表情参数）；$\boldsymbol{B}_{\mathrm{t}}$ 为纹理的 PCA 基；δ 为纹理 PCA 基对应的参数（简称纹理参数）。本文采用广泛应用的 BFM09[24]构建 $\overline{\boldsymbol{S}}$、$\boldsymbol{B}_{\mathrm{id}}$、$\overline{\boldsymbol{T}}$ 和 $\boldsymbol{B}_{\mathrm{t}}$，基于 FaceWarehouse[25]构建表情基 $\boldsymbol{B}_{\mathrm{exp}}$。

在模型训练阶段，给定一张红绿蓝三通道图像 $\boldsymbol{I}$，基于上述神经网络模型回归 3DMM 参数和相机参数，进行可微分三维人脸重建与渲染，得到重建图像 $\boldsymbol{I}'$，形式化表示如公式（3）～（4）所示。

$$\beta,\gamma,\delta,p,l=\mathrm{ResNet}(\boldsymbol{I}) \tag{3}$$

$$\boldsymbol{I}'=\mathrm{DifferentiableRendering}(\beta,\gamma,\delta,p,l) \tag{4}$$

其中，p 为姿态的参数；l 为光照的参数。

基于 $\boldsymbol{I}$ 和 $\boldsymbol{I}'$，本文分别采用图像级别的损失函数和感知级别的损失函数联合训练，共同对模型进行优化。图像级别损失函数包含像素损失函数和关键点损失函数。其中，像素损失函数如公式（5）所示。

$$L_{\mathrm{photo}}(\boldsymbol{I},\boldsymbol{I}')=\left\|\boldsymbol{I}'-\boldsymbol{I}\right\|_2 \tag{5}$$

即度量原输入图和重建图像之间的逐像素的差异。关键点损失函数通过常用的关键点检测算法为原图和重建图分别检测关键点$\{q_n\}$，计算公式如公式（6）所示。

$$L_{\mathrm{lan}}(\boldsymbol{I},\boldsymbol{I}')=\frac{1}{N}\sum_{n=1}^{N}\left\|q'_n-q_n\right\|_2 \tag{6}$$

其中，N 为关键点的总数，在本节中为 68。

除使用直观的图像级别损失函数外，本节还采用了感知损失函数。目前，较为流行的感知损失函数一般采用预训练的深度神经网络提取图像特征，并对图像特征之间的差异进行度量。

由于本节算法主要针对人脸，因此，采用大

规模预训练的人脸识别神经网络 FaceNet[26]提取人脸图像特征。具体表示如公式(7)所示。

$$L_{\text{per}}(\boldsymbol{I},\boldsymbol{I}')=1-\frac{\langle f(\boldsymbol{I})\cdot f(\boldsymbol{I}')\rangle}{\|f(\boldsymbol{I})\|\cdot\|f(\boldsymbol{I}')\|} \tag{7}$$

其中，$f(\cdot)$为提取图像特征的函数(神经网络)；$\langle\cdot\rangle$为向量内积。具体训练流程和数据集参考 Deng 等[12]的研究。

3DMM 参数回归模型将二维图像中的人脸投影到三维空间中，作为一种中间表示，其能够复现较多的人类面部表情。此外，3DMM 参数显式地将人脸形状解耦为身份、表情和姿态，解决了人脸属性耦合的问题。3DMM 参数回归模型是本文算法中的通用模块，训练完成后，若将该算法迁移到其他拓扑结构的三维虚拟人模型时，不需要再重新训练该部分。

3.1.2 blendshape 系数估计模型

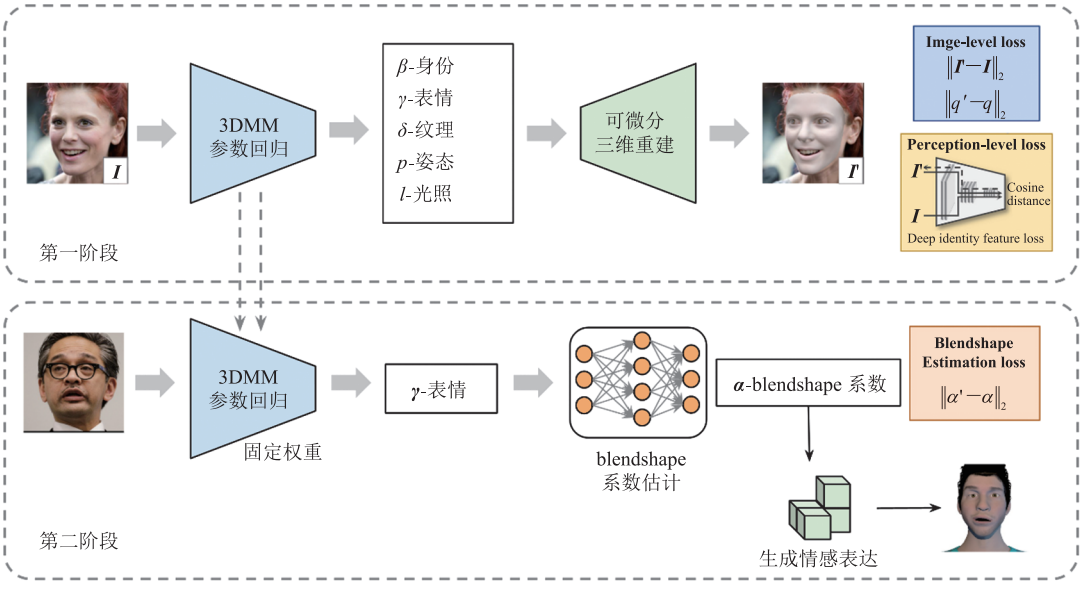
如图 1 所示，训练完成 3DMM 参数回归模型后，blendshape 系数估计模型将 3DMM 参数中的表情参数作为输入，使用神经网络预测目标三维虚拟人模型的 blendshape 系数。该模型是一个轻量化的神经网络模型，仅由两个线性全连接层组成，在两层之间有一个激活函数。在最后一个全连接层，Clamp 运算符将输出值截断为 0～1。在训练过程中，本文使用均方误差作为损失函数来训练优化该模型。

该神经网络模型的主要目的是构建 3DMM 表情参数和 blendshape 系数之间的映射关系。其中，3DMM 表情参数来源于 3DMM 中定义的 PCA 表情基，blendshape 系数则基于用户自定义的拓扑结构。根据 blendshape 拓扑结构的不同，该映射关系的建模复杂度也不同，在某些特定结构下，可能为二者建立简单的线性映射关系，而在相对复杂结构下，则需要更为复杂的非线性映射。因此，本文采用非线性的神经网络模型，理论上可增强通用性。

给定 3DMM 表情参数 γ 和真实的 blendshape 系数 α 标签，这一阶段可以形式化地表述为公式(8)～(9)。

$$\alpha'=T(\gamma) \tag{8}$$

$$L=\|\alpha'-\alpha\|_2 \tag{9}$$



注：该框架由两个阶段组成：(1)训练 3DMM 参数回归模型；(2)借助预训练的 3DMM 参数回归模型，训练 blendshape 系数估计模型

图 1　整体框架示意图

Fig. 1　Schematic diagram of the overall framework

其中，α' 为 blendshape 系数估计模型 T 预测的输出 blendshape 系数。模型训练的目标是最小化损失函数 L。

对于 blendshape 系数估计模型的训练，需要根据三维虚拟人模型的 blendshape 组织相应的数据集。本文借助三维建模软件 Maya 自动化构建数据集。如图 2 所示，给定目标三维虚拟人模型，所需的数据集包含：(1)随机生成的 blendshape 系数；(2)基于 blendshape 系数渲染生成的虚拟人面部图像。关于 blendshape 系数的生成，可将每个 blendshape 视为一个值为 0～1 的通道，系数生成的任务即为每个通道分配一个值。为避免极端的或不合理的面部表情，本文使用基于规则的方法生成 blendshape 系数。首先，为每个通道随机生成一个 0～1 之间的值；然后，使用一组基于 FACS 运动单元(action unit，AU)的规则，以剔除非法的 blendshape 值组合情况。例如，一个人几乎不可能同时向两个相反的方向移动他的嘴唇，若出现这种数值组合情况，那么应该剔除掉该组数值。由于诸如此类的规则约束存在，可以获得一套相对合理的 blendshape 系数值。关于虚拟人面部图像的生成，本文使用 Maya 软件中的渲染函数，基于生成的 blendshape 系数，渲染生成虚拟人脸图像，并使用一个正对虚拟人人脸的正面虚拟摄像头渲染虚拟人脸图像。

在得到目标数据集后，对 blendshape 系数估计模型的参数进行随机初始化，并将其置于预训练过的 3DMM 参数回归模型进行训练。首先，将数据集中的虚拟人脸图像输入 3DMM 参数回归模型，获取相应的 3DMM 参数；然后，将 3DMM 参数中的表情参数输入 blendshape 系数估计模型，令该模型输出预测的 blendshape 系数；最后，通过最小化预测的 blendshape 系数和数据集中真实的 blendshape 系数之间的误差，对 blendshape 系数估计模型进行优化。在该过程中，3DMM 参数回归模型的参数不再被训练优化，只有 blendshape 系数估计模型的参数被更新。

3.2 实现细节

对于 3DMM 参数回归模型的训练，本文使用公开的真实世界人脸数据(包括 CelebA 数据集和 LFW 数据集)。在数据准备阶段，本文利用 Deng 等[12]提供的方法对人脸图像进行人脸检测、对齐和裁剪，图像统一缩放为 224×224 的大小。在训练阶段，本文使用 ImageNet 数据集预训练的权重初始化 ResNet 神经网络，采用 Adam 优化器，设置 batch size 为 5 进行训练优化，训练学习率设置为 1e−5，共训练 300k 次迭代。

3DMM 参数回归模型是一个与 blendshape 拓扑结构无关的模块。经预训练后，对于不同

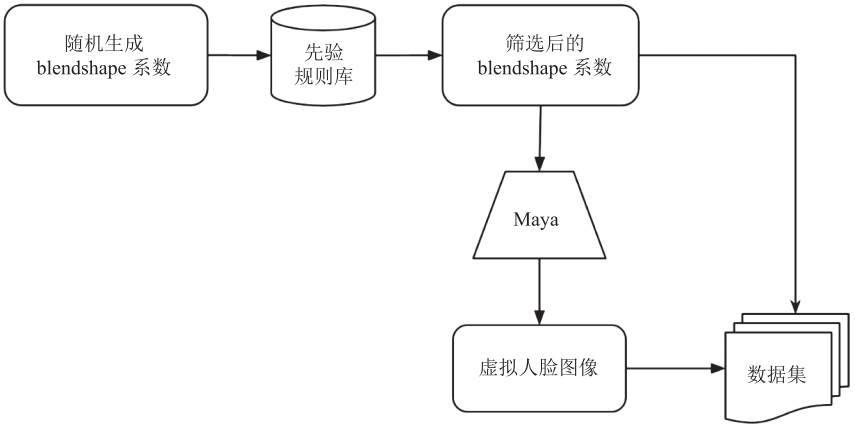


图 2 blendshape 数据集准备流程示意图

Fig. 2 Schematic diagram of blendshape dataset preparation process

blendshape 拓扑结构的虚拟人，该模块都无须进行任何额外的训练，即该模块是可复用的。

blendshape 系数估计模型是一个与虚拟人 blendshape 拓扑结构相关的模块。经训练后，该模型可适用于相同 blendshape 拓扑结构的虚拟人家族，即使他们的纹理外观不同。在实际应用中，经常会遇到 blendshape 拓扑结构不同的虚拟人模型。因此，本实验采用基于 FACS 规则制定的 50 个 blendshape。在其他研究项目中，可能涉及设计师按照个人喜好自行制作的 blendshape。在训练该 blendshape 系数估计模型时，需要根据目标的三维虚拟人模型不同的 blendshape 拓扑结构，相应地组织不同的数据集。然而，当面临新的 blendshape 拓扑结构时，可以自动构建一个新的数据集，然后对 blendshape 系数估计模型进行微调更新，并不影响本文方法的通用性和可迁移性。

4　评估与验证

上述两个模块经训练后，模型可以生成静态的虚拟人情感表达。此外，模型还支持生成动态情感表达，具体生成方式取决于源数据的类型。若源数据是图像，那么首先基于该图像估计 blendshape 系数，生成虚拟人面部表情，并将其作为峰值强度，然后通过对自然表情状态的 blendshape 系数值 0 到峰值强度 blendshape 系数值进行线性插值，获得平滑过渡的虚拟人动态情感表达。若源数据是视频，那么估计每一帧的 blendshape 系数，生成逐帧的虚拟人情感表达，然后将其拼接为动态情感表达。

本节基于一组虚拟人模型评估本文方法。该组模型共包括 6 个不同性别和外观的虚拟人模型，他们具有相同的 50 个 blendshape 拓扑结构，包括 46 个面部肌肉 blendshape 和 4 个眼睛注视 blendshape。面部肌肉 blendshape 的制作主要基于 FACS 的 46 个主要动作单元（AU1～AU46），用于刻画眉毛、眼睛、脸颊、嘴唇、嘴巴、鼻子、芯片、下巴等部位的运动；4 种眼睛注视的 blendshape 基于眼球运动动作单元（AU61～AU64），包括眼睛左转、右转、上转和下转。有关每个动作单元的详细说明，请参阅 Ekman 等[19]的面部动作编码系统。在本实验中，模型只需训练一次，即可实现在 6 个虚拟人中通用的效果。

4.1　blendshape 系数估计误差分析

在定量实验中，本文使用平均绝对误差（mean absolute error，MAE）直观地衡量该算法的准确性。MAE 公式定义如公式（10）所示。

$$MAE=\left\|\alpha'-\alpha\right\|_1 \tag{10}$$

其中，α 为数据对应的 blendshape 系数标签。

为找到合适的 blendshape 系数估计模型，本文对几种不同的设置进行了比较。

如表 1 所示，前 3 个实验对比了两个线性层之间不同隐层维度带来的不同误差效果。以 256 为基线模型，结果表明，较小的维度可能会限制模型的性能，将隐层维度增加到 384 时，没有观察到误差显著变化。然后，对不同的网络结构进行探索，使用 Leaky ReLU 函数替换 ReLU，在模型后附加 Clamp 运算符以及两个操作的组合。由表 1 可知，通过 Clamp 操作符截断输出范围，有利于减少估计误差，这是因为 blendshape 的合法系数值在 0～1 之间，截断操作可以强制性

表 1　blendshape 系数估计模型不同设置下的实验结果

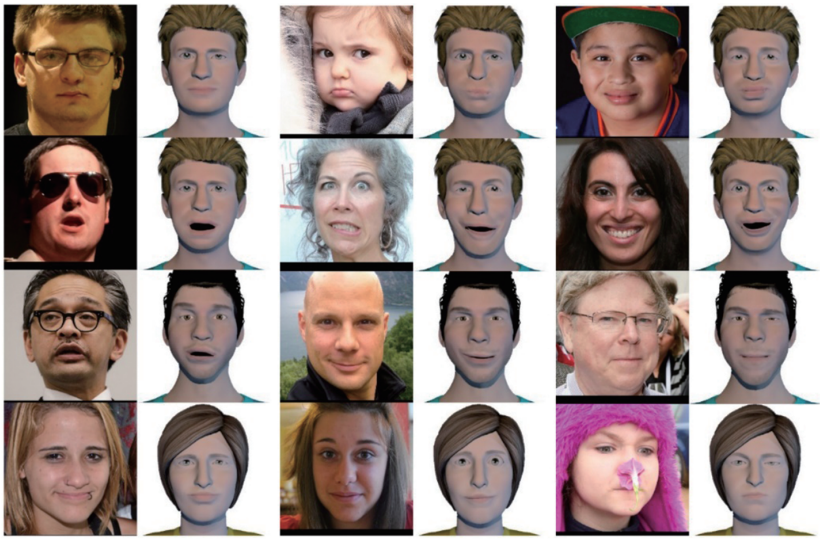
Table 1　Experimental results of the blendshape coefficient estimation model under different settings

Layers	Hidden-dim	MAE
Linear->ReLU->Linear	256	0.09
Linear->ReLU->Linear	100	0.10
Linear->ReLU->Linear	384	0.09
Linear->LeakyReLU->Linear	256	0.09
Linear->ReLU->Linear->Clamp	256	0.08
Linear->LeakyReLU->Linear->Clamp	256	0.07

过滤非法输出。实验结果表明，同时使用 Leaky ReLU 和 Clamp 可以使模型误差最小化。

4.2 虚拟人情感表达生成可视化结果

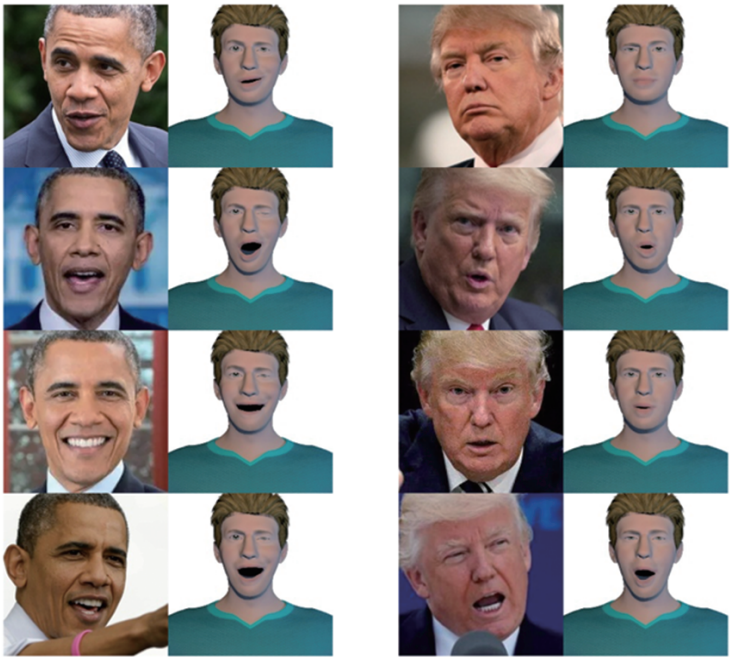
定性实验结果如图 3～4 所示，使用 blendshape 系数估计算法，基于真实的人脸图像创建虚拟人情感表达。图 3 展示了不同真人、不同虚拟人角色的表情重建结果；图 4 展示了对同一人物不同表情的重建结果。由图 3～4 可知，



注：每一列中，左侧为真实人脸表情图片，右侧为本文算法生成的虚拟人情感表达

图 3 blendshape 系数估计定性实验结果

Fig. 3 Qualitative experiment results of blendshape coefficient estimation



注：每一列中，左侧为真实人脸表情图片，右侧为本文算法生成的虚拟人情感表达

图 4 同一人物不同表情的重建效果

Fig. 4 Different expressions reconstruction of the same person

使用该算法创建的虚拟人面部表情能够生动地复现真实人脸图像的面部表情，虽然部分细微的肌肉运动无法重现，但是该算法自动估计的系数仍然可以提供一个合理的 blendshape 系数初始值。例如，图 3 中第 3 行第 1 列，真实人脸中嘴角微微向下的表情没有很好地被虚拟人脸复现，但眼睛、眉毛、唇部等部位都获得了较好的复现效果，研究者仅需在此基础上进行简单的微调，即可轻松地达到理想的效果。

4.3　与设计师的对比

将本文方法与设计师在虚拟人情感表达制作方面的表现进行对比，具体评价指标包括满意度得分和所用时长两个层面。满意度得分主要用于帮助用户度量虚拟人情感表达是否真实、自然，以及与真人情感表达是否接近，满意度分数范围为 1～10，分数越高表示满意程度越高。评估结果来自 14 名志愿者对 7 个随机挑选的样本进行评分。评估过程中，志愿者主要针对样本结果进行评分，对于情感表达的生成过程没有参与。实验结果如表 2 所示。

在满意度层面，尽管设计师相比于本文提出的方法获得了更高的满意度分数，但其差异并不显著。此外，根据标准差可以发现，志愿者打分的变化范围较大，这也侧面反映了人类评估情感表达时的主观性。

在耗时层面，首先观察到，生成或调整 blendshape 时，运行整个算法本文方法仅耗时 0.41 s，设计师则需要 236 s。巨大的时间差表明，本文方法可有效减轻设计师制作虚拟人情感表达的工作量，且能保持较高的逼真度和满意度。

对于本文提出的方法，初始化时间即训练 3DMM 参数回归模型和 blendshape 系数估计模型所需的时间；迁移 blendshape 时间意味着将该方法迁移至新的 blendshape 拓扑结构所需的时间，即重新训练 blendshape 系数估计模型的时间。对于设计师而言，初始化时间即设计师熟悉每一个 blendshape 所代表的形变意义所需时间；迁移 blendshape 时间即设计师熟悉一组新的 blendshape 拓扑结构所需时间。尽管在初始化时间和迁移 blendshape 时间层面，可以观察到本文算法较设计师需要更长的时间，但算法消耗的时间为机器时间，无须设计师干预，大幅降低了设计师的工作量。

表 2 最后一行呈现了，设计师在本文算法生成的 blendshape 结果基础上进行微调，设计师仅需约 1 min 即可完成 blendshape 微调，且取得了更高的满意度得分。由此可见，基于本文方法生成的情感表达亦可以作为设计师制作情感表达时的一个强有力助手。此外，针对新的 blendshape 拓扑结构，需要对本文模型的第二个模块进行自动化微调更新，从而完成模型迁移，耗时约 30 min。若模型为相同 blendshape 拓扑结构的虚拟人生成表情则不需要此步骤。

4.4　限制与不足

针对虚拟人情感表达工作，本文依然存在一定的限制与不足。首先，本文方法不能精细地复刻真人情感表达，在准确度上仍有一定提升空间；然后，使用本文提出的算法模型，对用户存在一定的学习曲线以及计算资源的要求；最后，尽管本文算法对于相同拓扑结构的三维虚拟人是

表 2　本文算法与设计师对比

Table 2　The algorithm in this paper is compared with the designer

	初始化时间	迁移 blendshape 时间	调整 blendshape 时间 (s)	满意度
本文算法	～hours	～30 min	0.41±0.06	6.36±1.80
设计师	199 s	199 s	236.12±48.61	6.92±1.91
本文算法＋设计师	—	—	71.31±15.22	7.21±1.86

完全通用的，但是面对新的拓扑结构，部分模型仍然需要重新训练。

5 结 论

针对虚拟人情感表达制作中耗时费力的问题，本文提出了一种基于人脸表情图像自动化估计 blendshape 系数的方法，实验与评估结果表明，本方法能够以较低的误差估计目标表情的 blendshape 系数，生成自然生动的虚拟人情感表达，进而有效减少设计师制作情感表达面部动画时的工作量。该方法通过引入 3DMM 参数回归模型，有效缓解了人脸图像中属性耦合的问题。针对具有相同 blendshape 拓扑结构的虚拟人，该方法可适用于任意不同的纹理，具有较强的通用性；针对不同的 blendshape 拓扑结构，该方法可以以较短的时间，自动化完成迁移适应，具有较强的迁移性。

参 考 文 献

[1] 张菁, 张天驰, 陈怀友. 虚拟现实技术及应用 [M]. 北京: 清华大学出版社, 2011.
Zhang J, Zhang TC, Chen HY. Virtual reality technology and its application [M]. Beijing: Tsinghua University Press, 2011.

[2] Caldas OI, Aviles OF, Rodriguez-Guerrero C. Effects of presence and challenge variations on emotional engagement in immersive virtual environments [J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2020, 28(5): 1109-1116.

[3] Volonte M, Hsu YC, Liu KY, et al. Effects of interacting with a crowd of emotional virtual humans on users' affective and non-verbal behaviors [C] // Proceedings of the 2020 IEEE Conference on Virtual Reality and 3D User Interfaces, 2020: 293-302.

[4] Kruzic CO, Kruzic D, Herrera F, et al. Facial expressions contribute more than body movements to conversational outcomes in avatar-mediated virtual environments [J]. Scientific Reports, 2020, 10(1): 1-23.

[5] Sylaiou S, Kasapakis V, Gavalas D, et al. Avatars as storytellers: affective narratives in virtual museums [J]. Personal and Ubiquitous Computing, 2020, 24(6): 829-841.

[6] Ekman P, Friesen WV. Constants across cultures in the face and emotion [J]. Journal of Personality and Social Psychology, 1971, 17(2): 124-129.

[7] Argyle M. Bodily communication [M]. London: Routledge, 2013.

[8] Bhattacharya U, Rewkowski N, Banerjee A, et al. Text2Gestures: a transformer-based network for generating emotive body gestures for virtual agents [C] // Proceedings of the 2021 IEEE Virtual Reality and 3D User Interfaces, 2021: 1-10.

[9] de Melo C, Paiva A. Expression of emotions in virtual humans using lights, shadows, composition and filters [C] // Proceedings of the International Conference on Affective Computing and Intelligent Interaction, 2007: 546-557.

[10] Blanz V, Vetter T. Face recognition based on fitting a 3D morphable model [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25(9): 1063-1074.

[11] Creswell A, White T, Dumoulin V, et al. Generative adversarial networks: an overview [J]. IEEE Signal Processing Magazine, 2018, 35(1): 53-65.

[12] Deng Y, Yang JL, Xu SC, et al. Accurate 3D face reconstruction with weakly-supervised learning: from single image to image set [C] // Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019: 285-295.

[13] Wolf L, Taigman Y, Polyak A. Unsupervised creation of parameterized avatars [C] // Proceedings of the IEEE International Conference on Computer Vision, 2017: 1530-1538.

[14] Shi TY, Yuan Y, Fan CJ, et al. Face-to-parameter

translation for game character auto-creation [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 161-170.

[15] Shi TY, Zuo ZX, Yuan Y, et al. Fast and robust face-to-parameter translation for game character auto-creation [C] // Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 1733-1740.

[16] Fourati N, Pelachaud C. Emilya: emotional body expression in daily actions database [C] // Proceedings of the 9th International Conference on Language Resources and Evaluation, 2014: 3486-3493.

[17] Niewiadomski R, Mancini M, Baur T, et al. Mmli: multimodal multiperson corpus of laughter in interaction [C] // Proceedings of the International Workshop on Human Behavior Understanding, 2013: 184-195.

[18] Ding Y, Prepin K, Huang J, et al. Laughter animation synthesis [C] // Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems, 2014: 773-780.

[19] Ekman P, Friesen WV. Facial action coding system [J]. Environmental Psychology & Non-verbal Behavior, 1978. https://doi.org/10.1037/t27734-000.

[20] Joshi P, Tien WC, Desbrun M, et al. Learning controls for blendshape-based realistic facial animation [M]. Berlin: Springer-Verlag, 2008: 162-174.

[21] Lewis JP, Anjyo K, Rhee T, et al. Practice and theory of blendshape facial models [J]. Eurographics (State of the Art Reports), 2014. http://dx.doi.org/10.2312/egst.20141042.

[22] Chuang E, Bregler C. Performance driven facial animation using blendshape interpolation [J]. Computer Science Technical Report, 2002, 2(2): 3.

[23] He KM, Zhang XY, Ren SQ, et al. Deep residual learning for image recognition [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[24] Paysan P, Knothe R, Amberg B, et al. A 3D face model for pose and illumination invariant face recognition [C] // Proceedings of the 2009 6th IEEE International Conference on Advanced Video and Signal Based Surveillance, 2009: 296-301.

[25] Cao C, Weng YL, Zhou S, et al. FaceWarehouse: a 3D facial expression database for visual computing [J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 20(3): 413-425.

[26] Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 815-823.