

基于少量图像的三维重建综述

于航¹, 付彦伟², 姜柏言³, 薛向阳^{1,2,3}

(1. 复旦大学工程与应用技术研究院, 上海 200082; 2. 复旦大学大数据学院, 上海 200082;
3. 复旦大学计算机科学技术学院, 上海 200082)

摘要: 基于少量图像的三维重建被认为是第三代人工智能的经典应用之一。在计算机图形学和计算机视觉领域, 基于少量图像的三维重建任务因具有广泛的应用场景和很高的研究价值, 长期以来吸引着众多学者的目光。引入深度学习方法后, 该领域于近年来得到了长足发展。对此类基于少量图像的三维重建任务进行了全面阐述, 并介绍了本研究组在该方面的系列工作, 对其中涉及的数据类型进行分析, 阐明其适用性和一般处理方法。此外, 对常见的数据集进行分析、整理, 针对不同重建方法, 归纳出其基本框架、思路。最后, 展示了一些常见三维重建的代表性实验结果, 并提出了未来可能的研究方向。

关键词: 三维重建; 小样本学习; 人体重建

中图分类号: TP391.41

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202240

A survey of image-based few-shot 3D reconstruction

YU Hang¹, FU Yanwei², JIANG Boyan³, XUE Xiangyang^{1,2,3}

1. Academy for Engineering & Technology, Fudan University, Shanghai 200082, China
2. School of Data Science, Fudan University, Shanghai 200082, China
3. School of Computer Science, Fudan University, Shanghai 200082, China

Abstract: Few-shot 3D reconstruction is considered one of the classic applications of the third generation of artificial intelligence. In the area of computer graphics and computer vision, few-shot 3D reconstruction has attracted the attention of many researchers during the past several decades because of its wide application scenarios and high research value. The area has grown significantly in recent years after the introduction of deep learning methods. The state-of-the-art methods in image-based few-shot 3D reconstruction were reviewed comprehensively and the series of works of our research group were introduced. The various 3D data types were introduced, and their applicability and general processing procedures in 3D reconstruction were discussed. Furthermore, the most widely used datasets were categorized. Finally, some representative experimental results of common 3D reconstructions were presented, and potential future research directions were proposed.

Key words: 3D reconstruction, few-shot learning, human reconstruction

0 引言

一般而言, 可以将三维重建问题看作照片成像的逆向问题。后者是把三维形状信息投影到二维图像空间, 而前者是根据二维照片信息, 估计出对应的三维形状。从单张图像习得的三维表示具有较大的模糊性, 即同一张图像, 理论上存在无数种可能

的三维结构。人类大多数时候会依靠过往的生活经验做出最可能的判断, 而人工智能系统却缺乏相应的“生活经验”, 正是这种模糊性导致该问题成为一个病态问题, 为了获得更准确的重建结构, 该领域既引入基于学习的计算机视觉方法, 同时也依赖计算机图形学的知识和数据结构。因此, 基于图像学习的三维重建任务属于计算机视觉、计算机图形

收稿日期: 2022-06-15; 修回日期: 2022-08-11

通信作者: 付彦伟, yanweifu@fudan.edu.cn

学和人工智能的交叉领域。

根据机器学习知识, 给定输入数据 $x \in X$ 和目标数据 $y \in Y$, 机器学习算法可以学习 3 种形式: 联合分布 $p(x, y)$ 、条件分布 $p(y|x)$ 、函数映射 $y = f(x)$ 。判别式任务常用第二和第三种形式; 生成式任务常用第一种形式, 比如生成式对抗网络 (generative adversarial network, GAN) 和变分自编码器 (variational autoencoder, VAE)。具体而言, 直接以生成式模型构建三维人体较为困难, 所以很多方法采用“曲线救国”策略: 首先利用函数将输入信息映射到特定低维编码空间, 再由对应解码器完成解码重建工作。所以, 在基于图像深度学习的三维重建任务中, 一般可采用第一种或第三种形式。

此外, 三维重建任务的评价与三维几何结构在世界坐标系中的位姿信息、相机参数无关。当物体的空间位置与理想的空间位置不同时, 可通过平移操作变换空间位置; 对于朝向, 也有旋转操作进行变换; 而尺寸则可随缩放改变。这些指标都不影响对重建效果的评价, 重建效果好坏只与物体表面外观是否准确有关。

基于图像的三维重建在学术研究和工业生产中都具有举足轻重的地位, 有着长久的研究历史。作为计算机图形学与视觉结合的重要课题, 基于少量图像样本的三维重建问题旨在根据给定的少量单视角或多视角图像, 估计其对应的三维信息。该任务从图像理解出发, 对图像特征进行感知与生成建模, 最终重建物体的三维模型。三维数据表示有体素、点云、隐函数、网格表示等。其中网格表示方式优点比较突出, 其可以保留物体表面细节, 而且计算量小, 易于纹理贴图, 在虚拟现实、增强现实等任务中应用比较广泛。

基于少量图像的三维重建任务本质上是极具挑战性的非适定病态问题 (ill-posed problem)。当解不唯一或定解条件不完整时, 如何从有限的二维图像输入恢复缺失的第三维度, 是考验神经网络学习、推理能力的问题。同时, 从图像推断三维形状是人类智能的基础能力, 对该任务的研究也可以验证基于深度学习的方法能否合理利用先验知识对三维空间进行感知与推断。故笔者认为基于少量图像的三维重建可以被认为是第三代人工智能的经典应用之一。因为该研究不仅结合人类对三维世界的知识与经验的推理, 还要依靠以深度学习模型为基础的数据驱动, 进而构成更强大的人工智能。

传统的基于图像的三维重建方法一般是基于多视角几何的方法, 利用标定的相机与视差较小的双目图像, 从几何角度用三角原理求解图像像素对应的三维坐标, 或者利用一系列连续图像帧, 联合估计相机的姿态变化与匹配特征点, 通过几何方式计算三维点云。但是这些依赖大量标定参数与图像序列, 因而对复杂的物体或场景适用性非常差。相较于传统的依赖光测度或几何方法的三维重建方法, 基于深度学习的三维重建具有对输入图像的依赖性弱、模型预测推理速度快、不需要光照和物体纹理限制等优点。

在深度学习蓬勃发展后, 基于深度学习的三维重建任务往往需要大量样本数据进行训练。但现实中, 既有真实照片又有以对应三维形状为训练真实值的样本数据过少。具体而言, 目前训练具有泛化能力的三维重建深度模型仍需要以大量标注好的图像和三维模型为训练数据, 这对于实际复杂场景建模任务来说是非常难的。学术界则较多利用 ShapeNet 等大规模仿真三维物体数据集^[1], 将图像渲染为训练数据。虽然合成数据有助于标注, 但合成数据与真实数据之间的图像域差异 (domain gap) 比较大。而真实的数据采集与标注格外昂贵, 所以大部分复杂动态物体或场景的三维重建问题只能获得少量训练样本。这就给一些在大量样本上可以实现良好效果的模型造成了困扰。因此, 对基于少量样本的三维重建任务进行深入研究具有突破性意义。现有的研究思路主要包括: 利用先验信息, 引入多种三维形状的初始模板进行形变; 将二维数据的方法迁移至三维数据空间中使用。人类天然具有想象物体三维形状的能力。得益于深度学习的强大表示能力, 基于深度学习的三维物体重建任务得以发展。基于深度学习的方法可以通过几张离散的图像恢复出物体的三维形状, 甚至只给定单张物体图像也能重建物体的大致三维形状。本文对相关代表性工作进行了详细的介绍。

一方面, 基于大样本强监督学习的模型预测能力有限, 难以获得“常识”, 很难满足开放环境中不断拓展的应用需求。另一方面, 当前主流深度学习模型依赖大量标签数据的强监督信息, 在少量样本条件下泛化性能有限乃至失效, 故基于少量样本监督信息的学习是人工智能待突破的瓶颈之一。本文将探讨并介绍基于少量图像的三维重建算法, 训练数据从经典的“大数据”到少量标注数据, 学习任务从二维图像空

间的识别与理解到三维空间的重建。

经典的“大数据+深度模型”学习模式在解决二维图像的理解上有很大优势,然而现实空间是三维的。三维视觉学习任务需要较好的建模空间先验知识,这也是人工智能研究中的一个重要问题。人类擅长从各种信号和模态中习得新的知识,并以特定形式表达出来。这是一个简单而又完整的学习过程,包含从既有信息中进行知识提取、建模,并依据习得的知识在新信息下做出推理、复现的过程。从艾伦·麦席森·图灵提出“机器能否思考”的问题开始,人工智能学科就为知识的推理建模而确立。在众多人工智能流派中,随着近10年计算机算力和训练数据量飞速发展,机器学习尤其是最近兴起的神经网络成了比符号主义更简单但效果更好的人工智能问题解决方案。依靠人类的先验知识,在推断时遵循奥卡姆剃刀定律,可以消除大部分由信息缺失带来的歧义,这一点在基于单张RGB图像的三维重建任务中尤为重要。因为单张RGB图像承载的信息本身具有较大模糊性,缺失了物体未呈现在镜头前的部分信息,人类只能依靠已有的经验推断。由于人类从小在环境中学习,这一推断过程在人类大脑中是自然而然的;但计算机的记忆中不存在这样“从小”学习的过程,所以在推断时需要为其提供额外信息。

除了针对静态物体的重建,基于深度学习的人体重建也是三维重建领域非常热门的研究方向之一,在骨骼驱动下,人体形状会发生显著变化,因此直接将针对通用物体重建的方法应用于人体重建往往无法取得令人满意的效果。为了解决这一问题,基于共享的骨架结构,多种人体参数化模型被提出,其中最具代表性的多人线性蒙皮模型(skinned multi-person linear model, SMPL)成为许多工作进行人体相关建模任务的基础,以实现从不同模态数据(包括彩色图像、视频以及点云等)进行人体重建。

从实际角度来看,基于深度学习的三维网格重建具有良好的应用价值,例如机器人对真实场景的理解,机械臂视觉抓取,人机在虚拟现实、增强现实层次的互动,医学影像诊断,三维打印,自动驾驶技术中的场景识别与重建,虚拟数字人等。但现有深度学习方法在泛化性、自适应性、可靠性上无法满足以上较复杂任务的应用要求。

从少量样本图像中重建出三维形状的任务,因其本身的特殊性与困难性,长久以来受到工业界和

学术界的持续关注。自深度学习技术发展以来,学术界从最初需要大量训练数据开始,如今已关注到样本量不足的重建任务研究。现有的三维物体重建任务主要基于ShapeNet数据集^[1],一般从中选取常见的十几种物体类别进行实验。同种类物体之间的几何结构相差不大。比如当需要重建的物体是飞机时,如果算法识别出图像中给出的是一架飞机,并对应输出一架飞机的通用几何结构,那么重建效果一般也不会太差^[2]。因此,该领域还有诸多问题有待进一步研究。

本文着重介绍三维重建涉及的数据类型、数据集、参数化模型、监督信息、评价指标和重建方法,并且对它们各自的适用性和性能进行分析。全文结构如下:第1节介绍三维物体重建工作,包括涉及的数据类型、方法与性能、数据集;第2节介绍与三维人体相关的一些典型应用,包括三维人体重建、姿态迁移、图像生成与四维人体建模;第3节总结全文并讨论一些开放性问题。

1 基于少量样本的三维物体重建

1.1 常用数据类型

1.1.1 RGB 图像

从输入图像或视频帧中提取特征的工作主要由深度卷积神经网络负责。通常情况下,需要将深度学习模型在ImageNet^[3]等大型图像数据集上做预训练。预训练后的模型可用于任何需要提取视觉特征的计算机视觉任务,如分类、检测、分割等。该类视觉信息编码器的设计是人工智能等领域的关键问题,所以本文不再赘述,请参考相关综述,如参考文献[4]。

1.1.2 视频(帧)

与RGB图像类似,视频数据也是人工智能模型中常用的输入信息。与光流^[5]类似,随着硬件设备的不断进步,拍摄视频也变得越来越容易,所以可以利用视频中记录的前后帧信息。因此,部分计算机视觉任务是专门针对视频类数据展开的。在使用视频数据时,一般需要根据不同任务的需求,提取特定帧,之后可以利用前后帧的关联信息,依靠图像处理的方式对提取的帧进行处理,并融合前后帧的上下文信息等。

1.1.3 视角信息

视角信息表示以相机(观察者)为中心建立坐标系,在相机坐标系下,物体的坐标即此时相机眼中它

的位置。每次对物体进行位置调整后,需要重新设定相机的外部参数矩阵。相机外部参数除了可以在制作数据集时自己设定,还可以利用运动推断结构算法(structure from motion, SfM)^[6]、基于视觉信息的同时定位与地图检测(visual simultaneous localization and mapping, VSLAM)算法进行估计。

1.1.4 三维模型

现有三维重建方法涉及多种三维模型,比如常见的CAD工程模型和常用的三维网格和体素表示法。其中以体素为形式的输出常受制于存储空间,无法实现精细的面部轮廓和肢体细节。高分辨率的体素表示带来了高昂的存储空间开销,使得现有的计算机很难对其进行存储和处理。当然,如今摩尔定律依然有效,体素表示的未来依然可期。有时为了与其他类型的数据进行比较,可能会将体素格式转换为网格等其他三维表示格式。主要使用的转换方法为行进立方体(marching cubes)算法^[7-8]。

另有隐函数(implicit function)类的三维表达形式可以精确建模三维形状。以符号距离函数(signed distance function, SDF)^[9]为代表的隐函数作为三维几何形体的表示方式,具有精确、计算量不大等优点。目前,隐函数作为人体表面边界点的有效约束,已经被越来越多地引入三维重建工作中。隐函数采样示意如图1所示,在特定采样空间内对三维网格进行采样,可得到内部点云和外部点云。

1.2 代表性方法

1.2.1 基于体素表示的方法

3D-R2N2^[10]是一种基于循环神经网络实现体素操作的三维重建模型。该模型从大量的合成数据中学习从物体图像到其基础三维形状的映射。该模型的网络结构从任意视角接收一个或多个物体实例的图像,并以三维占用网格的形式输出该物体的重建。与之前的大多数工作不同,3D-R2N2不需要任何图像注释或物体类别标签进行训练或测试。相关实验分析也表明,3D-R2N2 优于最先进的单视图

重建方法,并能够在传统的SfM/SLAM方法失败的情况下对物体进行三维重建。

1.2.2 基于点云表示的方法

(1) 点集生成网络

通过深度神经网络生成三维形状已经在研究界引起了越来越多的关注。大多数现存的工作采用常规的表示方法,如深度图或三维网格;然而,这些表示方法会丢失三维形状在几何变换下的部分细节,并且还存在着其他问题。点集生成网络(point set generation network, PSG)^[11]解决了从单一图像进行三维重建的问题,产生了一种直接的输出形式——点云坐标。PSG模型具有新颖而有效的架构、损失函数,并设计了一个点云采样组件,能够从输入图像中预测出多个可信的三维点云。在实验中,该模型不仅在基于单幅图像的三维重建基准上优于最先进的方法,而且在三维形状上也表现出很强的性能,在多种三维形状生成方面也有很好的泛化能力。

(2) PointNet

与卷积神经网络从二维图像中提取特征相似,PointNet^[12-13]可以从三维点云中提取形状的特征,为从物体分类、部分分割到场景语义解析等应用提供了一个统一的架构。PointNet直接将点云作为输入,输出整个输入的分类标签或输入的每个点的分段/部分标签。PointNet及其改进版本PointNet++在当前涉及点云处理的工作中发挥着重要作用,它们凭借轻量的结构取得了非常好的实验效果,因而被广泛采用。

上述方法都为基于点云表示的三维重建任务提供了非常值得借鉴的解决思路,点云虽然是一种非常高效的三维表示方式,但是其只能通过空间中离散点来表示三维形状,点与点之间相互独立,无法提供更丰富的表面信息。

1.2.3 基于网格表示的方法

(1) Pixel2Mesh

Pixel2Mesh(以下简称P2M)是利用三维形状的先验知识完成少量样本三维重建的代表性工作

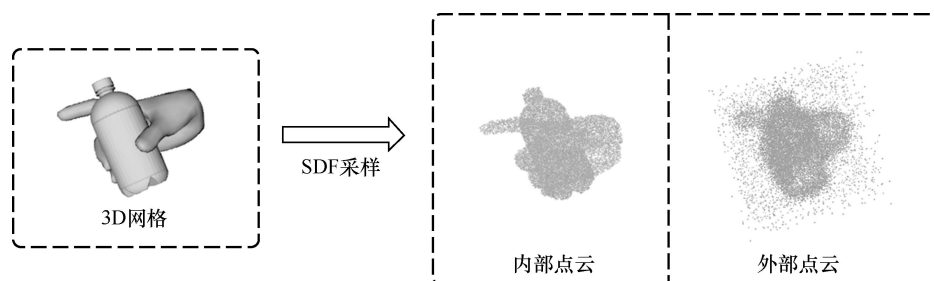


图1 隐函数采样示意

之一。如图2所示, P2M^[14]是一个端到端的深度学习架构, 可以从单一的彩色图像中生成一个三角形的三维形状。受限于深度神经网络的性质, 以前的方法通常使用体素或点云来表示三维形状, 而将它们转换为具有更精细表面结构的三维网格并不容易。与现有的方法不同, P2M在图卷积神经网络中表示三维网格, 并利用从输入图像中提取的感知特征, 对椭球网格进行形变, 产生正确的几何形状。同时, P2M采用从粗到细(coarse-to-fine)的策略使整个形变过程稳定。此外, 通过定义各种与网格有关的损失, P2M可以提取不同层次的属性, 以保证三维几何形状视觉上美观、结构上准确。广泛的实验表明, 该模型不仅能生成具有更高细粒度的网格模型, 而且与最先进的方法相比, 其实现了更高

的三维形状估计精确度。P2M能够基于单视图实现完整的三维形状重建, 核心在于其利用精巧设计的特征提取和网格形变神经网络(网格变形网络), 从大量数据中学习到了不同类别的形状先验, 在测试时能够根据输入图像判定其类别并以从粗到细的方式不断使重建变得精细。另外, 基于网格定义的一系列损失函数起到了关键的正则化作用, 确保形变过程相对平滑, 在一定程度上缓解了网格自交叉的问题。

(2) Pixel2Mesh++

Pixel2Mesh++(以下简称P2M++)模型结构如图3所示, P2M++^[15]在P2M的基础上, 研究了从少量带有相机姿势的彩色图像中生成三维网格表示的重建问题, 其中相机姿势可以由用户提供, 也可以通过网络进行预测。虽然许多以前的工作直接

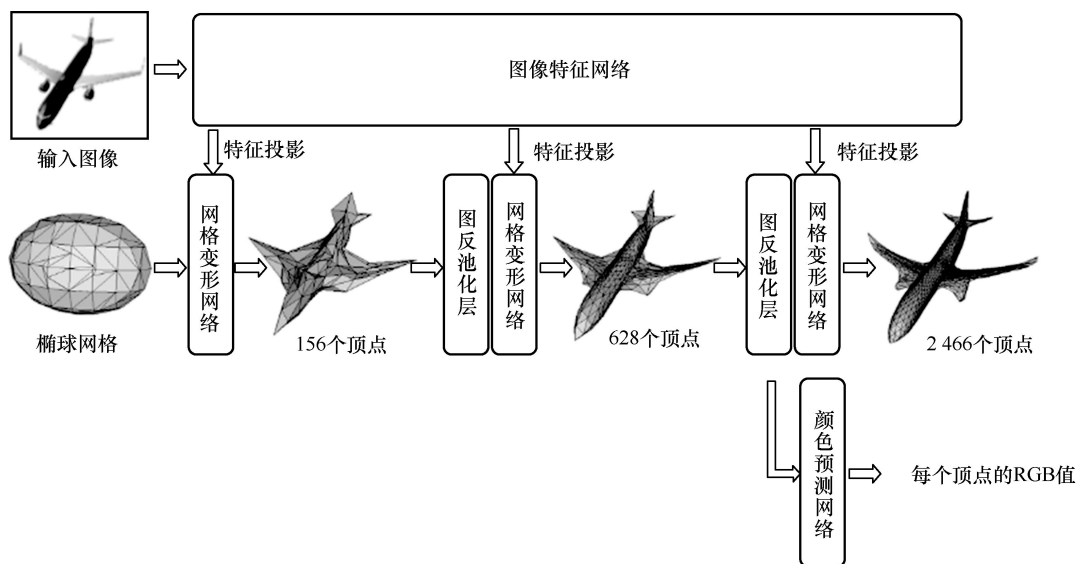


图2 P2M模型结构^[14]

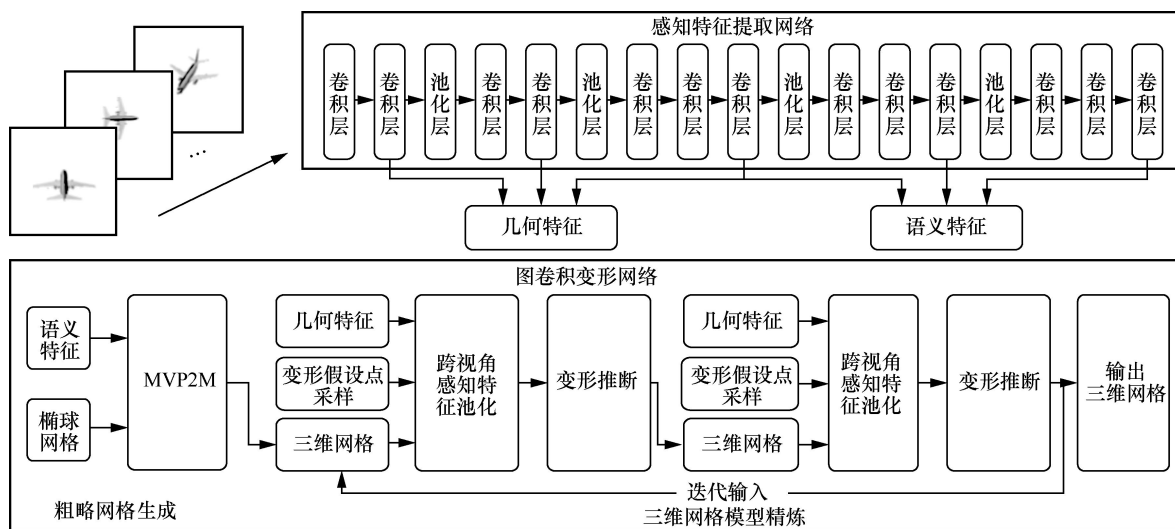


图3 P2M++模型结构^[15]

从先验条件中学习、幻化出形状,但 P2M++通过图卷积网络的跨视角信息进一步提高形状质量。与 P2M 类似,该模型不是建立一个从图像到三维形状的直接映射函数,而是预测每个网格顶点的偏移量,以迭代地改进一个粗略的形状。受传统多视图几何方法的启发, P2M++的网络针对初始网格顶点位置附近的区域构建一个正十二面体,并从中进行采样,利用从多个输入图像中建立的感知特征统计来推断最佳变形, P2M++形变采样点示意如图 4 所示^[15]。广泛的实验表明,该模型生成了准确的三维形状,不仅从输入的角度看是可信的,还能很好地与任意视点对齐。在物理驱动架构的帮助下,该模型还表现出了跨越不同语义类别和输入图像数量的概括能力。模型分析实验表明, P2M++对初始网格的质量和相机姿势的误差具有鲁棒性,并且可以与可微渲染器结合起来优化测试时间。

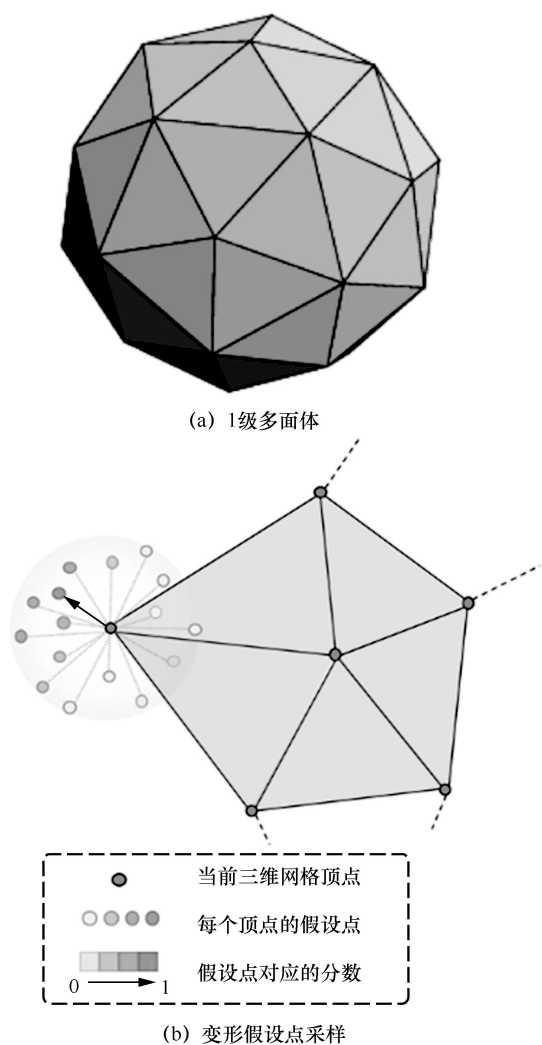


图4 P2M++形变采样点示意^[15]

(3) 神经三维网格渲染器

传统图形学渲染管线中的光栅化操作是一个离散化的过程,无法支持梯度的反向传导。神经三维网格渲染器(neural 3D mesh renderer, N3MR)^[16]是一种可微分的渲染器,对光栅化过程产生的梯度进行近似估计,将神经网络生成的三维网格模型以可微分的形式渲染为二维图像,在训练时利用二维轮廓信息进行监督以支持端到端的训练。N3MR 在三维重建任务中可以不依赖真实的三维数据,仅使用二维图像进行训练,并且它能够作为通用组件与基于网格表示的三维重建方法结合,在缺少三维监督信息的情况下使训练成为可能。

(4) AtlasNet

AtlasNet^[17]是一种利用参数化表面生成三维形状的表面的方法。AtlasNet 将三维形状表示为参数化表面元素的集合,与生成体素网格或点云的方法相比,AtlasNet 可以自然地推断出形状的表面表示。除此之外,AtlasNet 还具有较好的泛化能力,以及生成任意分辨率形状的可能性,而且没有内存问题。这些优点使得 AtlasNet 在 ShapeNet 等数据集上可以更好地完成变形、参数化、超级分辨率、匹配和共同分割等任务。

1.2.4 基于隐函数的方法

(1) OccupancyNet

OccupancyNet^[18]是一种基于学习的三维重建方法的新表示方法。它利用隐式表达将三维表面表示为一个深度神经网络分类器的连续决策边界。与现有的方法相比,OccupancyNet 的表示方法以无限的分辨率对三维输出的描述进行编码,而没有过多的内存占用。相关实验也验证了这种表示方式可以有效地编码三维结构,并且可以从各种输入中直接推断出完整的三维形状。进一步的实验表明,对于在单一图像、噪声点云和粗糙的离散体素网格中进行三维重建的挑战性任务,OccupancyNet 在定量和定性结果上都具有竞争力。因此有理由相信,OccupancyNet 将成为各种基于学习的三维任务的有用工具。基于隐式曲面的表示方式近年来逐渐成为重建任务中最热门的选择,甚至有研究者将其应用于图像超分辨等任务中,并取得了令人印象深刻的结果。隐式曲面的主要优势在于:①定义简洁,非常适合用深度神经网络进行建模;②相比基于网格形变的方法需要借助许多相对复杂的损失函数来约束训练过程,隐函数只需要简单的分类/回归损失以及针对隐空间的正则化约束即可进行稳定训练。但是,基于真实

网格表面计算用于训练的 Occupancy/SDF 值是一个计算代价很大的过程, 因此相比其他三维表示方式, 基于隐函数的方法在数据预处理阶段更加耗时。

(2) DeepSDF

DeepSDF^[9]是一类形状的连续有符号距离函数(signed distance function, SDF)表示法, 能够从部分和嘈杂的三维输入数据中获得高质量的形状表示细节。DeepSDF 和其他利用隐式表达的三维重建模型相同, 都是使用一个连续的隐式空间来表示形状的表面: 该空间中一个点的值表示该点到表面边界的距离, 符号表示其是在形状的内部(正号)还是外部(负号), 因此 SDF 的零等值面即物体的表面边界, 同时也明确地将空间分割为物体表面内部与外部。此外, 该模型在学习三维形状表示方面显示了较为优越的性能, 同时与以前的工作相比, 模型的大小减少了一个数量级。DeepSDF 除了在重建任务中取得了非常惊艳的效果, 其另一重要贡献是提出了自动解码(autodecoding)的模式, 即不需要编码器, 直接基于

观测数据对隐变量进行反向优化, 在节省了显存消耗的同时还能够得到更精细的重建结果。

1.3 代表性数据集

三维物体数据集如图 5 所示, 自深度学习被广泛应用以来, 与三维重建任务相关的数据集得到了充分利用, 并促进了相关研究迅猛发展。

1.3.1 Pascal3D+

Pascal3D+^[19]由真实拍摄的图像组成, 包含多种场景和拍摄模式。Pascal3D+ 包含从 PASCAL VOC 2012 数据集中选出的 12 类刚性物体。这些物体都有位姿信息的标注(方位角、仰角和与摄像机的距离)。Pascal3D+ 还添加了来自 ImageNet 数据集的 12 类物体的位姿标注图像。

1.3.2 Pix3D

Pix3D^[20]是一个大规模的三维数据集, 其中包括各种图像与形状的二维-三维像素级对齐信息。Pix3D 在与形状相关的任务中有着广泛的应用, 包括重建、检索、视线估计等。



图5 三维物体数据集

1.3.3 ObjectNet3D

ObjectNet3D^[21]是一个用于三维物体识别的大型数据集，它由100个类别、90 127张图像组成，同时包含这些图像中的201 888个物体种类和44 147个三维形状。制作该数据集需要首先将图像中的物体与模型库中的三维形状对齐，然后为每张图像提供最接近的三维形状标注和准确的三维位姿标注。因此，该数据集有助于从二维图像中识别物体的三维姿态和三维形状。

1.3.4 ShapeNet

ShapeNet^[1]是一个具有丰富注释的大规模三维形状数据集，由普林斯顿大学、斯坦福大学和丰田工业大学芝加哥分校的研究人员合作完成。它为世界各地的研究人员提供数据，以促进对计算机图形学、计算机视觉、机器人学和其他相关学科的研究。该数据集包含大量CAD模型及其纹理信息。ShapeNet包含以下子集。①ShapeNetCore。ShapeNetCore包括单一的三维模型以及人工验证的类别和对齐注释，它涵盖了55个常见的物体类别和大约51 300个独特的三维CAD模型。②ShapeNetSem。ShapeNetSem是一个更小、更密集的有标注子集，包括12 000个模型，分布在更广泛的270个类别中，除了人工验证的类别标签和一致的对齐方式，这些模型还标注了真实世界的尺寸、类别层面的材料组成估计以及总体积和重量的估计。

1.3.5 KITTI

KITTI^[22]是用于移动机器人和自动驾驶的非常

受欢迎的数据集之一，涵盖了用各种传感器记录的数小时的交通情况：包括高分辨率的RGB、灰度立体相机和三维激光扫描仪。尽管它很受欢迎，但该数据集本身并不包含用于语义分割的标注。然而，不同的研究人员已经对该数据集的部分内容进行了手工注释，以适应他们的需要。

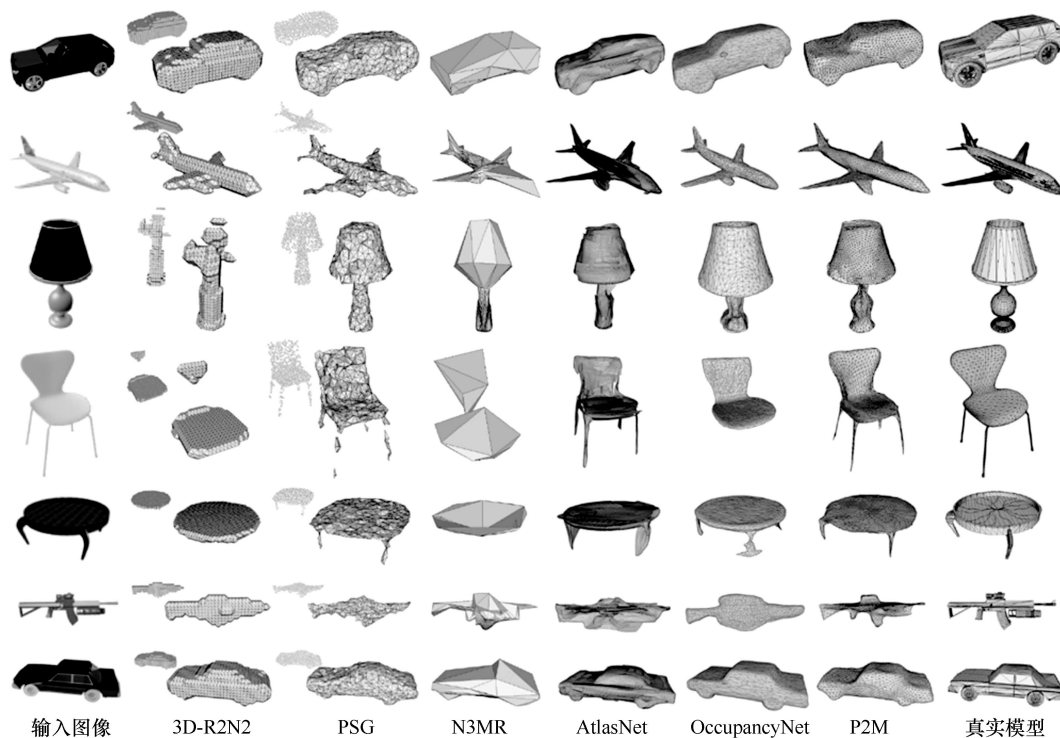
1.4 主要实验结果

表1展示了主要的三维重建模型在ShapeNet数据集特定子类别上的定量测量结果。其中，CD表示倒角距离(chamfer distance)，EMD表示动土距离(earth mover distance)，均越小越好。每种类别下最好的模型结果用粗体字表示。从表1可以看出，P2M在大部分类别下有较好的结果，并取得了最佳的平均准确度。PSG在一些类别上的表现也较为出色。它依靠三维点云生成三维形状。点云表示法在展示几何细节方面有更大的灵活性和能力，但如果不进行适当的后处理，不一定能生成更好的网格模型。

图6展示了前文提及的主要方法在ShapeNet数据集上的定性实验结果。从图6可以看出，由于分辨率低，3D-R2N2产生的三维体素结果缺乏细节，例如图6第4行所示，在椅子的例子中缺少腿，另外如果提高其输出的分辨率还会增加大量的存储开销。PSG产生了稀疏的三维点云，CD误差相对较小。然而，由于缺乏与表面有关的损失函数(如表面法线)，这些点云不一定能产生平整的三维网格。对于基于三维网格类的方法，N3MR产生的形状非常粗糙，这对于少数渲染任务来说可能满足需

表1 定量实验结果(结果来自P2M^[14])

类别	CD					EMD				
	3D-R2N2	PSG	N3MR	AtlasNet	P2M	3D-R2N2	PSG	N3MR	AtlasNet	P2M
飞机	0.895	0.430	0.450	0.468	0.477	0.606	0.396	7.498	0.659	0.579
长凳	1.891	0.629	2.268	0.703	0.624	1.136	1.113	11.766	1.204	0.965
衣橱	0.735	0.439	2.555	0.433	0.381	2.520	2.986	17.062	2.503	2.563
汽车	0.845	0.333	2.298	0.340	0.268	1.670	1.747	11.641	1.407	1.297
椅子	1.432	0.645	2.084	0.724	0.610	1.466	1.946	11.809	1.534	1.399
显示器	1.707	0.722	3.111	0.848	0.755	1.667	1.891	14.097	1.641	1.536
台灯	4.009	1.193	3.013	1.575	1.295	1.424	1.222	14.741	1.350	1.314
音响	1.507	0.756	3.343	0.812	0.739	2.732	3.490	16.720	3.108	2.951
枪支	0.993	0.423	2.641	0.461	0.453	0.688	0.397	11.889	0.735	0.667
沙发	1.135	0.549	3.512	0.621	0.490	2.114	2.207	14.876	1.910	1.642
桌子	1.116	0.517	2.383	0.577	0.498	1.641	2.121	12.842	1.673	1.480
手机	1.137	0.438	4.366	0.443	0.421	0.912	1.019	17.649	0.868	0.724
船只	1.215	0.633	2.154	0.839	0.670	0.935	0.945	11.425	0.907	0.814
平均	1.432	0.593	2.629	0.680	0.591	1.501	1.652	13.386	1.500	1.380

图6 定性实验结果^[14]

求,但不能恢复出复杂物体的形状,如椅子和桌子等。AtlasNet 使用 25 个小块(patch)和 2 500 个点进行测试,它的结果相对增加了更多的细节,但仍然包含强烈的噪点和孔洞。OccupancyNet 是这个领域中较为先进的模型之一。可以看出,OccupancyNet 可以产生具有光滑表面细节的三维网格,例如图 6 中的汽车和飞机,但偶尔会忽略物体的一些小部分,例如椅子的腿。P2M 在大部分类别下也有准确的形状。虽然产生了相当好的结果,但当目标几何体不能用亏格(genus)为 0 的网格表示时,P2M 通常会补充网格空洞,使拓扑结构亏格为 0,如图 6 中的机枪。

2 基于少量样本的三维人体重建

作为三维重建任务的一个主要子类别任务,三维人体重建有着广泛的应用场景,比如在线会议、虚拟试装等。但该任务与三维物体重建的思路略微不同。人体的某些部位极其细致,如手指尖和面部五官,少量点凸起或凹陷,表面就会呈高低起伏的锯齿状。直接由二维图像信息映射至三维几何空间的重建方法效果并不稳定^[11-18],因此目前大多以人体参数化模型辅助重建。从数学角度来讲,人体参数化模型本质是对顶点坐标信息进行降维,压缩人体重建的解空间。假设原有网格具有 N 个顶点,需

学习 $3N$ 个维度的数据。但使用人体参数化模型后,避免了位移、尺寸和朝向等因素对最终重建效果的影响,只需学习低维的姿态和体型等信息。可以将利用人体参数化模型学习三维人体重建看作在一个约束条件下求极值的问题。

2.1 少量样本的三维人体重建典型应用举例

针对少量图像的三维建模问题,本节进一步介绍如下几个在实际问题中的研究工作,涉及三维人体重建、神经姿态迁移以及四维人体建模等问题。

2.1.1 三维人体重建

HMR^[23]以单张 RGB 图像为输入数据,基于弱透视投影模型,直接回归出相机参数以及人体的形状与姿态参数,除了以往常用的二维关节点监督信息,还增加了三维监督和对抗网络监督,显著提升了其效果。

HMD^[24]主要基于从图像中提取的二维监督信息,对 HMR 方法生成的初始人体网格进行形变。但其最终效果往往不如 HMR,还会产生诸如三维信息丢失或生理结构错乱等问题。

CMR(convolutional mesh regression)^[25]将 ResNet 作为图像特征提取器,同时构建了一个基于图卷积神经网络的网格变形网络。工作时以 SMPL 参数化模型的初始人体为模板,将图像特征传给网格变形网络,从初始人体模板开始变形,利用点对点的监督等信息使初始模板最终变成目标网格的形状。在实践

中,CMR 在面比较极端的人体姿态或拍摄角度时,表现并不稳定。其还有一个经过多层感知机回归成 SMPL 模型的版本,输出更加稳定

PIFu^[26]不仅基于体素,还利用了隐函数和网格等表达来辅助完成重建任务。PIFu 利用多层感知机来提取图像特征,利用隐函数判断顶点在表面内还是外,还利用另一个隐函数回归顶点的颜色,即 RGB 值。另外,该模型可以读入多视角数据。为了追求理想效果,作者最后还利用行进立方体算法^[7-8],将生成的人体体素转换成人体网格。因为之前学习过表面纹理信息,所以转换后的人体网格是带纹理的。在实践中,基于体素的方法会丢失某些人体部位的细节信息,比如面部细节。而 PIFu 更多依赖纹理学习来给人体加上细节。

HMV^[27]方法基于 SMPL 参数化模型,输入多

视角的人体图像,引入二维和三维的监督信息,进而重建出姿态较为准确、体型非常平滑饱满的人体。其他方法^[28]则是 P2M^[14]方法在少量样本三维人体建模上的直接扩展。

本文简单对比了多个三维人体重建方法的结果,如图 7 所示。从图 7 可以看出,基于 SMPL 参数化模型的 HMR 和 HMV 都具有相对饱满和平滑的身体形状;而基于三维网格的 HMD 和 CMR 出现体型略微不平整的情况;基于隐函数的 PIFu 因采样空间的缘故,出现较多人工痕迹(artifact)。整体而言,该方向依然值得继续研究。

2.1.2 神经姿态迁移问题

姿态迁移方面已经有几十年的研究基础,姿态迁移旨在将原始姿态应用于目标身体模板上。神经姿势迁移(neural pose transfer, NPT)示意^[29]

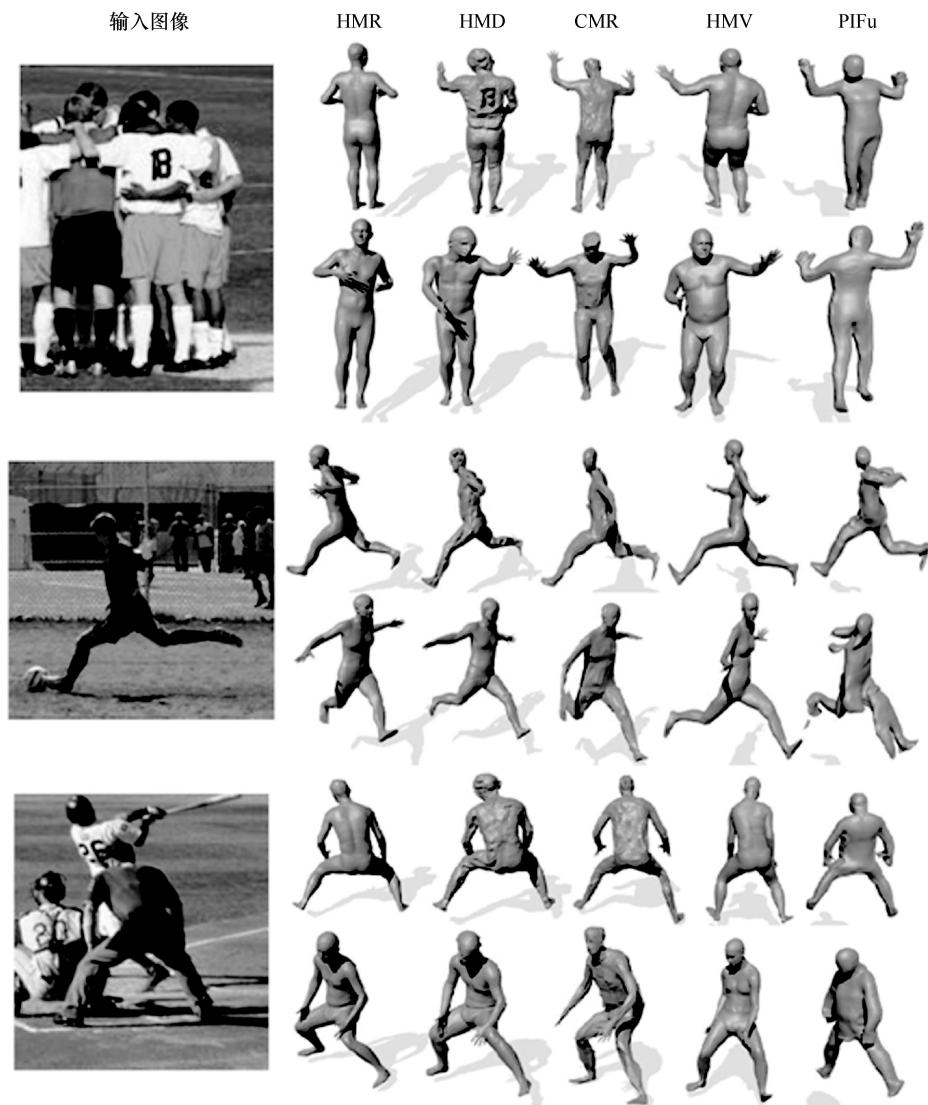


图7 不同三维人体重建方法结果

如图8所示,该项工作是基于少量样本完成人体姿态迁移的代表性工作,属于将二维图像风格迁移思路应用于三维数据的典型代表。作者将源人体网格的姿态迁移到目标人体网格的变形上,而源人体网格和目标人体网格可能来自不同的人物体型。以往的研究一方面需要成对的源人体网格和目标人体网格之间包含点对点的对应关系,可能需要大量的人工标注。另一方面,当源人体网格和目标人体网格具有不同特征时,深度模型的泛化能力是有限的。为了打破这一限制,作者提出了第一个神经人体姿态迁移模型,通过最新的图像风格转移技术解决姿势转移问题,利用新提出的组件——空间自适应实例规范化。该模型不要求源人体网格和目标人体网格之间有任何对应关系。相关实验表明,该模型能够有效地将变形从源人体网格迁移到目标人体网格,并具有良好的泛化能力。

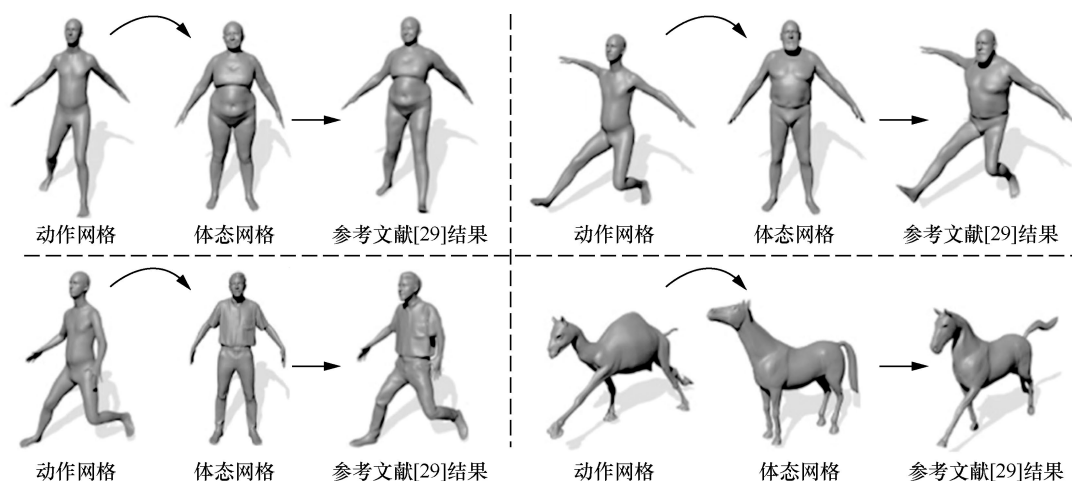


图8 神经姿势迁移示意^[29]

2.1.3 人体姿态引导图像生成

在人体姿势的引导下生成逼真的图像是一项有挑战性的任务。尽管以前的工作在合成标志性视图中的图像方面取得了成功,但在其他视图中的姿势引导的图像合成任务方面没有太多先例。人体姿态引导任务图像生成示意^[30]如图9所示,在此项工作中,作者提出了一个新的框架多分支细化网络MR-Net^[30],它对目标人物姿势、前景人物身体和场景图像等视觉线索进行解析,辅助图像生成过程。此外,作者还提出了一种新的兴趣区域(region of interest, RoI)感知损失来优化MR-Net。在两个非标志性数据集 Penn Action 和 BBC-Pose 以及一个标志性数据集 Market-1501 上进行了广泛的实验。实验结果表明,该模型能够解决姿势引导的人物图像问题。

2.1.4 四维人体建模

尽管基于深度学习的三维重建取得了令人印象深刻的结果,但直接学习具有精细几何形状的四

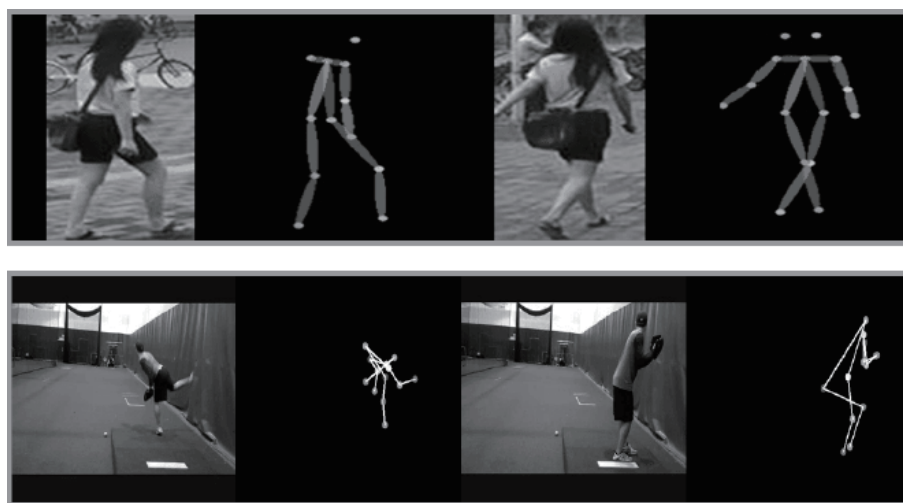


图9 人体姿态引导任务图像生成示意^[30]

维人类捕捉模型的技术却较少。这里的四维指三维空间以及时间维度。

(1) 基于神经常微分方程的方法

基于学习的表示已经成为许多计算机视觉系统成功的关键。虽然已经提出了许多三维表示法, 但如何表示动态变化的三维物体仍然是一个未解决的问题。基于神经常微分方程的四维人体建模示意^[31]如图 10 所示, 作者提出了一种深度学习模型对四维对象(在一个时间跨度上变形的三维对象)进行组合表示, 将对象的形状、初始姿态和运动信息解耦, 每个组成部分都通过一个低维的隐向量来表示。为了建立运动模型, 作者训练了一个神经常微分方程(neural ordinary differential equation, neural ODE)^[32]来更新初始状态, 并以学到的运动编码指导解码器利用形状编码和更新的姿态编码来重建每个时间戳的三维模型。此外, 作者还提出了一种身份交换训练策略, 以鼓励网络有效地学习正确解耦形状和姿态信息。相关实验表明, 该方法在四维重建方面优于现有的最先进的基于深度学习的方法, 并在各种与四维场景有关的任务上有明显的提升。

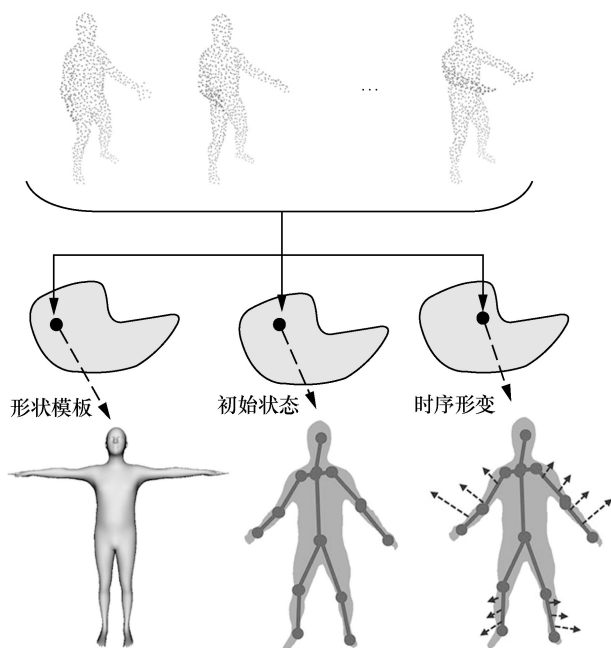


图 10 基于神经常微分方程的四维人体建模示意^[31]

(2) 基于参数化模型的方法

H4D 示意^[33]如图 11 所示, 这项工作提出了一个新的框架, 通过广泛使用的 SMPL 参数化模型中提供的人体先验, 可以有效地学习对动态人类进行紧凑的组合表示。作者提出的表示方法名为 H4D, 通过

SMPL 的形状和初始姿态参数, 以及编码运动和辅助信息的隐向量, 在一个时间跨度上表示动态的三维人体。作者提出了一个简单而有效的线性运动模型, 以提供一个粗略的、规范化的运动估计, 然后利用辅助编码中的信息对姿态和几何细节进行逐帧补偿。在技术上, 作者引入了新的基于门控循环单元(gated recurrent units, GRU)^[34]的架构, 以促进学习并提高表示能力。相关实验表明, 该方法不仅在恢复具有准确运动和精细几何形状的动态人体方面具有功效, 而且适用于各种与四维人体相关的任务, 包括运动重定位、运动补全和未来运动预测。

2.2 数据集与评价指标

2.2.1 代表性数据集

自深度学习蓬勃发展以来, 具有矫正标注并且数据量充足的三维数据集对三维模型的训练起到至关重要的作用。三维人体重建发展时间线如图 12 所示, 其中也展示了部分具有代表性的数据集。常用的三维人体数据集信息见表 2。数据集的类型代表该数据集是通过真实人体采集的(真实), 还是利用参数化模型人工生成的(人工)。数量指提供的数据条目数, 如果数据集已预先划分好训练集、验证集、测试集, 则它们的数量用“/”分别表示。如没有公开可查询的具体数字, 则数据后带有 K/M 用以表示数量级, $K = 10^3$, $M = 10^6$ 。模特数表示其中为了采集使用的真实人物数量, 更多模特的主要贡献在于增加数据集中体型的多样性。如果是彩色图像数据集, 则姿态数指代该数据集中涵盖的人体姿态种类; 如果数据集提供的是视频(帧), 则其场景数指模特按规定动作完成的场景数量。提出机构表示提出该数据集的大学或科研院所。MPI 是马克斯·普朗克研究所(Max Planck institute), IMAR 是罗马尼亚科学院数学研究所(institute of mathematics of the Romanian academy), Technion 是以色列理工学院(Israel institute of technology), IRI.UPC-CSIC 是西班牙国家研究委员会机器人与工业信息研究所(the institut de robòtica i informàtica industrial of the Spanish national research council and the technical university of catalonia)。开源性指该数据集的获取方式: “公开”表示在其官方网站可免费下載; “申请”表示申请者需要填写申请表并交由官方部门审核, 审核通过后为申请者发放下载链接或许可; “注册”表示需要在其官方网站注册, 然后方可下載; “私有”表示非直接共享, 若想获取需要单独与作者联系。

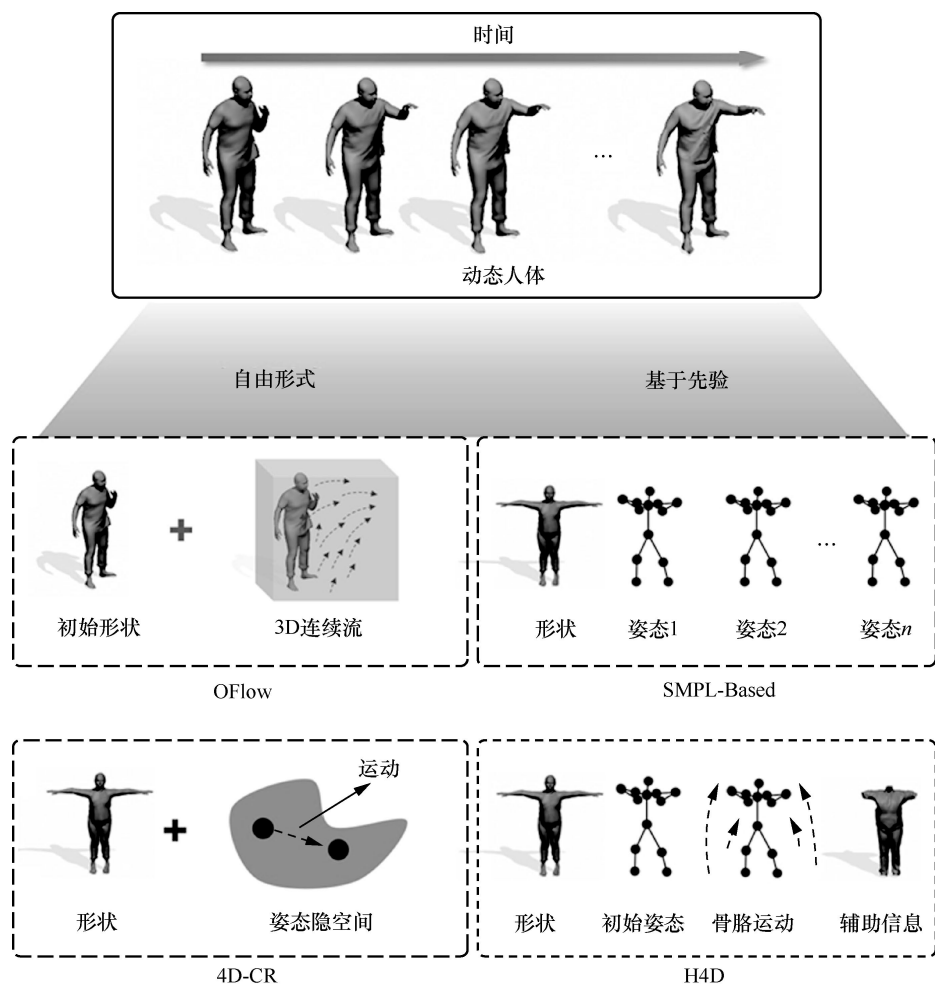
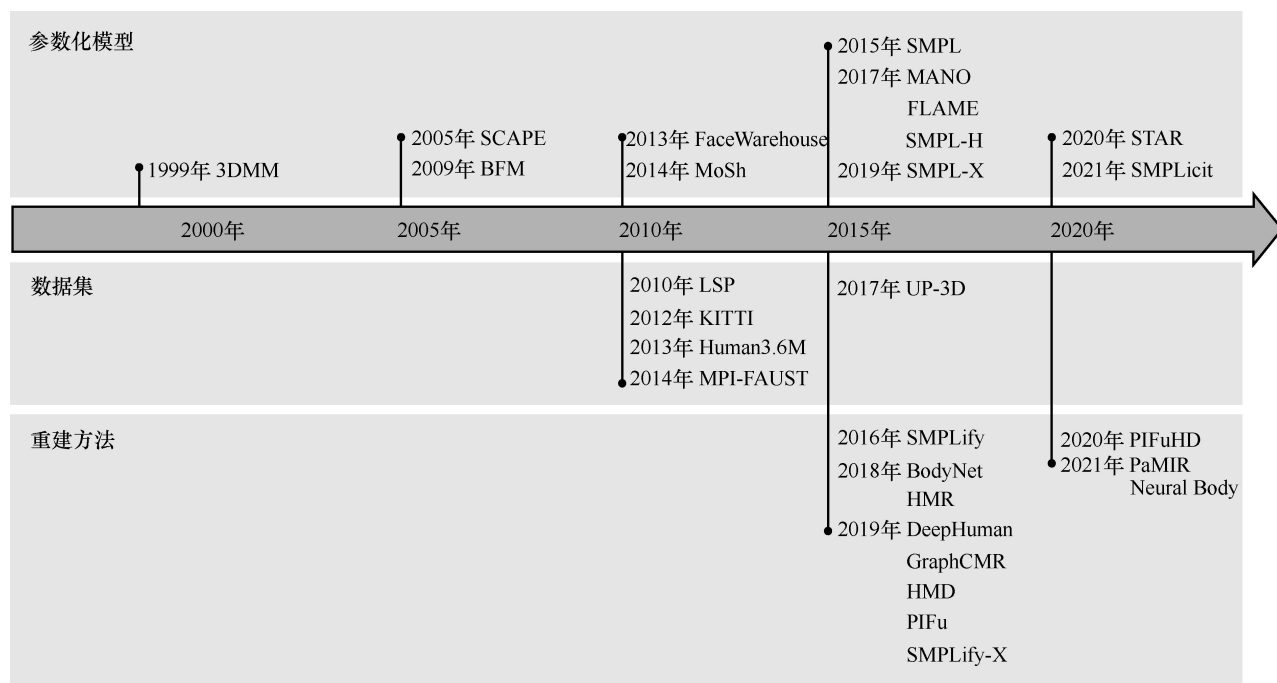
图 11 H4D 示意^[33]

图 12 三维人体重建发展时间线

表2 常用的三维人体数据集信息

数据集(缩写)	类型	数量	模特数	姿态数/场景数	提出机构	开源性
LSP	真实	2K	2K	2K	利兹大学	公开
LSP-extended	真实	10K	10K	10K	利兹大学	公开
HumanEva-I	真实	37K/6.8K/24K	4	76	MPI/布朗大学	注册
HumanEva-II	真实	2 460	2	2	MPI/布朗大学	注册
Human3.6M	真实	3.6M	11/3/4	15	IMAR	申请
TOSCA	人工	39	27/12	5	Technion	公开
SHREC	人工	1 184	1 184	5	Technion	公开
FAUST	真实	200/100	10	1	MPI	注册
DFAUST	真实	40K	10	1	MPI	注册
BUFF	真实	11 054	6	11 054	MPI	注册
MPII-Pose	真实	25K	40K	25K/410	MPI	注册
MS-COCO	真实	250K	250K	250K	MPI	公开
SURREAL	人工	5 901/5 281/2 118	115/30	1 964/703	MPI	申请
UP-3D	人工	8 515	5 703/1 423	8 515	MPI	公开
UP-S31	人工	8 515	8 515	8 515	MPI	公开
UP-P14	人工	8 128	8 128	8 128	MPI	公开
UP-P91	人工	8 128	8 128	8 128	MPI	公开
AMASS	真实	11K	300	11K	MPI	注册
FHB	真实	100K	26	45	MPI	申请
ObMan	人工	21K	2 772	21K	MPI	申请
RECON	人工	150	150	150	南京大学	公开
SYN	人工	300	300	300	南京大学	公开
Lifescan	真实-人工	—	2 043	2 043	MPI	私有
3DPeople	真实	2M	80	70	IRIUPC-CSIC	公开
People Snapshot	人工	自提取帧	24	11	MPI	申请
MINI	人工	12K	12	12K	MPI	申请
Multi Garment	人工	96	96	96	MPI	注册
THUman	真实	7K	7K	7K	清华大学	申请
THUman2.0	真实	500	50	500	清华大学	申请
PROX	真实	100K	20	12	MPI	注册
GRAB	真实	1.6M	5/5/51	1 334	MPI	注册
HUMBI	真实	1.5K	772	772	MPI	注册

虽然近年来三维人体数据集已经得到了极大丰富和完善,但依然存在缺乏统一认可的数据集的问题。一个重要原因是,该类数据集涉及人体隐私及版权,采集和发布存在较大困难。在早期三维人体重建的工作中,研究者基本使用以前的人体姿态估计所用的数据集,比如 LSP^[35]、HumanEva^[36]、Human3.6M^[37]等。

2.2.2 常用评价指标

作为三维重建的一个分支,第一节中的 CD 与 EMD 同样可以作为三维人体重建精度的评价指标。除此之外,还有一些基于三维人体估计和重建设计

的常用评价指标。

如果只关注估计人体的三维姿态,MPJPE (mean per joint position error) 是当前应用最广泛的评价指标,其测量预测关节点和真实关节点之间的平均欧氏距离,数值越小表示估计的准确性越高。在 MPJPE 的基础上,研究者提出了 PA-MPJPE (Procrustes aligned-MPJPE),通过普罗克鲁斯特斯 (Procrustes) 对齐方法来消除平移、旋转和尺度的影响,从而仅专注于三维骨架的准确性,该数值同样是越小越好。此外,PCK (percentage of correct keypoints) 通过人工设置的阈值来判定关键点估计

是否正确,并计算了正确估计出的关键点的比例,该指标数值越大越好。

如果除了姿态参数还估计了 SMPL 形状参数,并且数据集中有对应的真实 SMPL 网格模型,则可以利用网格表面的拓扑结构一致性,采用 PVE (per vertex error) 来计算预测的网格顶点和真实网格顶点之间的平均欧氏距离,数值越小越好。

3 结束语

本文介绍了基于少量图像的三维重建这一当今计算机视觉和计算机图形学领域中的热门研究方向的发展历史、主要方法、数据集,并简述各种数据类型特点,对部分常见模型的性能做出了分析。

但该领域还存在很多开放性问题值得探究,在此简要讨论。

① 实时性重建:实时性在数字孪生等应用方向有着重要作用。但其计算量较大,目前四维人体重建方法的推断过程较少实现实时性^[38]。

② 对抗样本干扰:可以在图像提取特征过程中加入干扰噪声,使判别器无法正确理解提取的特征。该方向在现今人体重建领域研究较少,但对于信息安全、人体重建工业化都至关重要^[39]。因此,该方向也将是未来研究的重点之一。

③ 大型真实数据集:与二维图像和视频数据处理的方法类似,三维重建也需要大量来自真实世界的不同类别的三维物体数据为训练样本,由于三维物体的捕获和扫描往往需要高精度的设备和配置好的场景,采集难度较大,因此目前仍然缺少大型的真实世界三维物体数据集。

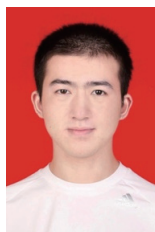
本文提到的姿态迁移、人体姿态引导图像生成、四维人体建模等一系列应用,在动画制作、机械臂抓取等相关工业场景有着重要的工程应用价值。

参考文献:

- [1] CHANG A X, FUNKHOUSER T, GUIBAS L, et al. ShapeNet: an information-rich 3D model repository[J]. arXiv preprint, 2015, arXiv:1512.03012.
- [2] TATARCHENKO M, RICHTER S R, RANFTL R, et al. What do single-view 3D reconstruction networks learn?[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 3400-3409.
- [3] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2009: 248-255.
- [4] WANG W, YANG Y J, WANG X, et al. Development of convolutional neural network and its application in image classification: a survey[J]. Optical Engineering, 2019, 58: 040901.
- [5] BARRON J L, FLEET D J, BEAUCHEMIN S S. Performance of optical flow techniques[J]. International Journal of Computer Vision, 1994, 12(1): 43-77.
- [6] ULLMAN S. The interpretation of structure from motion[J]. The Interpretation of Visual Motion, 1979: 133-175.
- [7] LORENSEN W E, CLINE H E. Marching cubes: a high resolution 3D surface construction algorithm[J]. ACM SIGGRAPH Computer Graphics, 1987, 21(4): 163-169.
- [8] LEWINER T, LOPES H, VIEIRA A W, et al. Efficient implementation of marching cubes' cases with topological guarantees[J]. Journal of Graphics Tools, 2003, 8(2): 1-15.
- [9] PARK J J, FLORENCE P, STRAUB J, et al. DeepSDF: learning continuous signed distance functions for shape representation[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 165-174.
- [10] CHOY C B, XU D F, GWAK J, et al. 3D-R2N2: an unified approach for single and multi-view 3D object reconstruction[C]//Proceedings of 2016 European Conference on Computer Vision – ECCV 2016. Cham: Springer, 2016: 628-644.
- [11] FAN H Q, SU H, GUIBAS L. A point set generation network for 3D object reconstruction from a single image[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 2463-2471.
- [12] CHARLES R Q, HAO S, MO K C, et al. PointNet: deep learning on point sets for 3D classification and segmentation[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 77-85.
- [13] QI C R, YI L, SU H, et al. PointNet++: deep hierarchical feature learning on point sets in a metric space[J]. arXiv preprint, 2017, arXiv:1706.02413.
- [14] WANG N Y, ZHANG Y D, LI Z W, et al. Pixel2Mesh: generating 3D mesh models from single RGB images[C]//Proceedings of 2018 European Conference on Computer Vision – ECCV 2018. Cham: Springer, 2018: 55-71.
- [15] WEN C, ZHANG Y D, LI Z W, et al. Pixel2Mesh++: multi-view 3D mesh generation via deformation[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 1042-1051.
- [16] KATO H, USHIKU Y, HARADA T. Neural 3D mesh renderer[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 3907-3916.
- [17] GROUEIX T, FISHER M, KIM V G, et al. A papier-mache approach to learning 3D surface generation[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 216-224.
- [18] MESCHEDER L, OECHSLE M, NIEMEYER M, et al. Occupancy networks: learning 3D reconstruction in function space[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 4455-4465.

- [19] XIANG Y, MOTTAGHI R, SAVARESE S. Beyond PASCAL: a benchmark for 3D object detection in the wild[C]//Proceedings of IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE Press, 2014: 75-82.
- [20] SUN X Y, WU J J, ZHANG X M, et al. Pix3D: dataset and methods for single-image 3D shape modeling[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 2974-2983.
- [21] XIANG Y, KIM W, CHEN W, et al. ObjectNet3D: a large scale database for 3D object recognition[C]//Proceedings of 2016 European Conference on Computer Vision – ECCV 2016. Cham: Springer, 2016: 160-176.
- [22] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2012: 3354-3361.
- [23] KANAZAWA A, BLACK M J, JACOBS D W, et al. End-to-end recovery of human shape and pose[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7122-7131.
- [24] ZHU H, ZUO X X, WANG S, et al. Detailed human shape estimation from a single image by hierarchical mesh deformation[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 4486-4495.
- [25] KOLOTOUROS N, PAVLAKOS G, DANIILIDIS K. Convolutional mesh regression for single-image human shape reconstruction[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 4496-4505.
- [26] SAITO S, HUANG Z, NATSUME R, et al. PIFu: pixel-aligned implicit function for high-resolution clothed human digitization[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 2304-2314.
- [27] LIANG J B, LIN M. Shape-aware human pose and shape reconstruction using multi-view images[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 4351-4361.
- [28] YU H, CHEANG C, FU Y W, et al. Multi-view shape generation for 3D human-like body[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2022.
- [29] WANG J S, WEN C, FU Y W, et al. Neural pose transfer by spatially adaptive instance normalization[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 5830-5838.
- [30] XU C M, FU Y W, WEN C, et al. Pose-guided person image synthesis in the non-iconic views[J]. IEEE Transactions on Image Processing, 2020, 29: 9060-9072.
- [31] JIANG B Y, ZHANG Y D, WEI X K, et al. Learning compositional representation for 4D captures with neural ODE[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 5336-5346.
- [32] CHEN R T Q, RUBANOVA Y, BETTENCOURT J, et al. Neural ordinary differential equations[J]. arXiv preprint, 2018, arxiv:1806.07366.
- [33] JIANG B Y, ZHANG Y D, WEI X K, et al. H4D: human 4D modeling by learning neural compositional representation[J]. arXiv preprint, 2022, arXiv:2203.01247.
- [34] CHO K, MERRIENBOER B V, BAHDAU D, et al. On the properties of neural machine translation: encoder-decoder approaches[C]//Proceedings of the 8th Workshop on Syntax, Semantics and Structure in Statistical Translation. [S.l.:s.n.], 2014: 103-111.
- [35] JOHNSON S, EVERINGHAM M. Clustered pose and nonlinear appearance models for human pose estimation[C]//Proceedings of the British Machine Vision Conference 2010. [S.l.:s.n.], 2010.
- [36] SIGAL L, BALAN A O, BLACK M J. HumanEva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion[J]. International Journal of Computer Vision, 2009, 87(1/2): 4-27.
- [37] IONESCU C, PAPAVA D, OLARU V, et al. Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(7): 1325-1339.
- [38] GUO K W, XU F, YU T, et al. Real-time geometry, albedo and motion reconstruction using a single RGBD camera[J]. ACM Transactions on Graphics, 2017, 36(4).
- [39] XIAO C W, YANG D W, LI B, et al. MeshAdv: adversarial meshes for visual recognition[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 6891-6900.

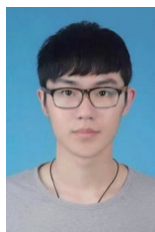
[作者简介]



于航(1991-), 男, 复旦大学工程与应用技术研究院博士生, 主要研究方向为 3D 视觉、人体重建。



付彦伟(1986-), 男, 博士, 复旦大学大数据学院研究员、博士生导师, 主要研究方向为少量样本学习、稀疏化学习算法、图像修复等。



姜柏言(1996-), 男, 复旦大学计算机科学技术学院博士生, 主要研究方向为三维/四维人体表示与重建。



薛向阳(1968-), 男, 博士, 复旦大学计算机科学技术学院教授、博士生导师, 主要研究方向为计算机视觉、多媒体内容识别和类脑智能系统等。