文章编号：2095-2163(2023)11-0232-07 中图分类号：TP183 文献标志码：A

基于深度学习的视觉手势估计综述

武 胜，秦浩东
（中国电子科技南湖研究院，浙江 嘉兴 314001）

摘 要：基于深度学习的视觉手势估计一直是计算机视觉领域的重点研究课题之一，随着深度学习和神经网络相关研究取得了巨大进步，针对手势估计中的高自由度、肤色、环境干扰、遮挡等问题已经远远优于传统方法。基于深度学习的三维手势估计主要是通过构建神经网络，对图像特征进行抽象化分析和理解，从而预测出手指关键点的三维坐标以及角度等信息，进而构建出手掌模型。准确的三维手势估计可以快速推动 AR/VR 行业的发展，因为沉浸与交互是 AR/VR 的关键要素，通过视觉手势交互可以为用户提供更方便、快捷、逼真的 AR/VR 互动体验。本文首先对当前手势估计方案进行阐述，了解到手势估计各方案的优缺点，然后介绍了基于深度学习的手势估计方法、相关数据集和评价指标，最后根据各研究结果，对当前三维手势估计所面临的挑战以及未来发展进行阐述。

关键词：手势估计；深度学习；关键点检测；神经网络

Overview of visual gesture estimation based on depth learning

WU Sheng，QIN Haodong
（China Nanhu Academy of Electronics and Information Technology，Jiaxing Zhejiang 314001，China）

Abstract：Vision gesture estimation based on deep learning has always been one of the key research topics in the field of computer vision. With the great progress of deep learning and neural network related research，it has been far superior to traditional methods for the problems of high degree of freedom，skin color，environment interference and occlusion in gesture estimation . Three-dimensional gesture estimation based on deep learning is mainly to construct a neural network for abstract analysis and understanding of image features，so as to predict the three-dimensional coordinates and angles of key points of fingers and then build a palm model. Accurate 3D gesture estimation can rapidly promote the development of AR/VR industry，because immersion and interaction are the key elements of AR/VR. Through visual gesture interaction，users can provide more convenient，fast and realistic AR/VR interactive experience. In this paper，firstly，the current gesture estimation schemes are described，and the advantages and disadvantages of each gesture estimation scheme are understood. Then，the gesture estimation method based on deep learning is introduced，and the related data sets and evaluation indexes are introduced. Finally，the challenges and future development of current 3D gesture estimation are described according to the research results.

Key words：gesture estimation；deep learning；key point detection；neural network

0 引 言

三维手势姿态估计是从采集的图像或者视频等对象中预测出手部关键点的位置[1]，再根据手关节点的位置预测出手掌的姿态，主要包含了目标识别、分割、回归检测等。传统手势估计受光线环境、拍摄角度、遮挡等影响，其准确性与实时性受到限制。随着卷积神经网络、递归神经网络、生成对抗网络等深度学习网络模型[2]的发展，以及 GPU 算力的提升，深度学习在图像分割、图像识别、图像分类方面已经取得了巨大进步，手势估计使用深度卷积神经网络，预测得将更加准确。目前，基于深度学习的研究方法基本可以划分为 3 类，分别是：基于点云的深度神经网络、基于体素的深度神经网络以及基于多视点的深度神经网络。

另外，随着计算机图形学、计算机视觉、人工智能等多学科的快速发展，苹果、谷歌、华为、微软等也都推出了相关的 AR/VR 引擎，AR/VR 相关成果已广泛应用于教育、医疗、军事等领域。虚拟与现实的

作者简介：秦浩东(2000-)，男，硕士研究生，主要研究方向：AR 软件开发、手势识别。
通讯作者：武 胜(1991-)，男，高级工程师，主要研究方向：AR 软件研究设计、unity3D 渲染引擎和手势识别研究。Email：wusheng@ mail.ustc.edu.cn

收稿日期：2022-11-14

交互是增强现实中不可或缺的一部分，手势交互[3]仍然是 AR/VR 最重要的交互方式，可以增强用户的沉浸感，利用手势可以实现远程操作、手语识别等应用，这也推动着视觉手势估计的进一步的发展。

本文主要对三维手势姿态估计进行梳理与分析，阐述基于深度学习的手势估计方法，整理相关数据集与评价指标，并对当前所面临的问题和未来发展趋势进行了阐述。

1　手势估计相关工作

1.1　手势估计方案分类

手势估计可分为 3 类：基于可穿戴设备的手势估计、基于深度传感器的手势追踪估计、基于视觉的手势估计。

(1)可穿戴设备的数据手套[4]通过内置传感器采集手部的运动数据，主要包括惯性、光纤以及光学三种传感器技术数据手套。基于惯性的数据手套虽然价格便宜，但是其漂移问题较为严重。基于光学的数据手套通过多个红外等摄像头采集手部数据，一般具有价格昂贵、遮挡等一系列问题。基于光纤的数据手套的数据精度以及稳定性虽然较好，但是其价格也十分昂贵，容易损坏。通常长时间穿戴数据手套存在手部会发汗，影响操作的沉浸感等问题，因此，数据手套没有得到大规模的应用。

(2)基于深度传感器的手势追踪估计[5]，如：Leap Motion 和 Kinect，在内部已经封装好手部重要信息识别算法，使用比较简单方便，但是其采集识别准确性取决于摄像机方向，这会限制用户的运动，而且在背景复杂、遮挡以及光线变化较大时，识别率较低。

(3)基于图像视觉的手势估计[6-7]可以解决价格昂贵、穿戴不方便等问题，但是仍然深受遮挡、光线等问题困扰，而就目前图像学、人工智能等学科的快速发展，基于视觉的手势识别仍然是研究的主流方向。基于视觉的研究方法可以分为基于双目的方法和基于 RGB 的方法以及基于 RGB-D 的方法。带有双摄像头以及深度传感器手机的普及，给视觉手势提供了条件。基于 RGB-D 的深度图与彩色图融合的方法有着其它方法所不具备的优势：

① 使用单一的深度图在超过一定距离后会出现精度下降情况，而彩色图相机具有变焦功能，可以容易获取较远距离的物体。

② 三维信息转换到二维信息过程中必将丢失一些数据，丢失的数据可以经过彩色图予以找回。

③ 单一的彩色图在计算深度数据上精度会出现误差，通过深度图可进行补偿计算。手势姿态估计方案如图 1 所示。

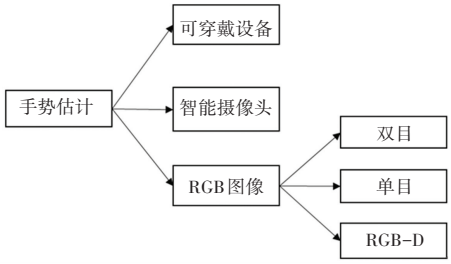

图 1　手势姿态估计

Fig. 1　Gesture pose estimation

1.2　手势运动学分析

手部由手指、手掌以及手腕共有 27 个互相连接的骨骼组成，手势估计最核心的问题是对手腕以及手指指骨的关节、连同指尖处进行识别、分割、跟踪以及估计，人手骨骼分布如图 2 所示。

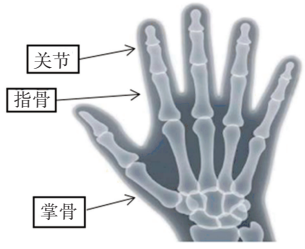

图 2　人手骨骼分布

Fig. 2　Distribution of human hand bones

人手是一个具有 26 自由度的执行机构，具体包括指骨关节 1 个弯曲自由度；掌骨关节 1 个自由度弯曲，1 个自由度绕转，故 2 个自由度；腕骨为 6 自由度，因此共有 1 * 2 * 5+2 * 5+6 = 26 个自由度，手掌 26 自由度模型如图 3 所示。

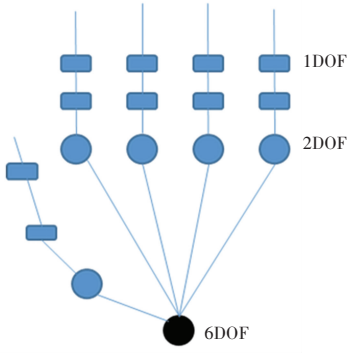

图 3　手掌 26 自由度模型

Fig. 3　26 degree of freedom model of the palm

根据人手指骨骼关节、手掌模型以及运动分析可以得出手部参与交互的主要为手指关节、掌指关节以及手腕[8]。因此，目前主流的手掌模型关节编

码有 14、16、21 三种，大多数论文以及数据集都是采用 21 关节点模型，通过估计关节点在三维空间的坐标，可预测出手姿态。手掌不同自由度模型如图 4 所示。

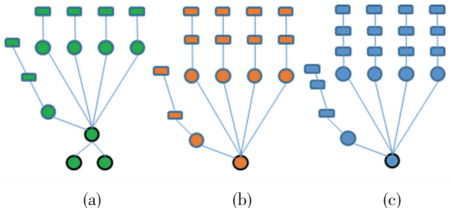

图 4 手掌不同自由度模型

Fig. 4 Models of the palm with different degrees of freedom

1.3 识别流程

手势估计包括人手识别、分割、跟踪、估计四步。其中，人手识别是为了减少背景噪声对手势估计的影响以及降低后续处理的计算量，识别出手部的区域。人手分割是将手部数据进行像素级别的提取，获取手部精准的信息。手部跟踪是通过连续帧预测下一步的手部位置，减少手部定位的耗时。手势估计是从图像中回归出手部完整的姿态，最终获取关节点三维坐标信息。

2 深度学习的手势估计方法

基于视觉的三维手势估计自首次引入深度学习以后，深度学习已经成为视觉手势的一个主流研究领域，越来越多的科研学者通过训练大量的样本数据，强化了模型的性能，获得了更加精准的特征，提高了鲁棒性以及泛化能力。基于深度学习的视觉估计可分为基于人工的神经网络、图神经网络、卷积神经网络、深度神经网络等[9-10]。根据 Erol 等学者[11]的综述结论，三维手势跟踪算法可以分为判别法、生成法[12]，而为了利用二者的优点，有学者提出了混合法。

2.1 判别法

判别法又称为数据驱动，对数据特别依赖，需要多个高质量的数据集，可学习从图像特征空间到手势特征空间的映射关系，进而预测出手势。判别法根据手势跟踪的检测与估计进行区分，又可以分为基于回归的方法与基于检测的方法。判别法由于可以采用离线的训练，无需大量手掌模型，因此，更适合实时应用。

2014 年，Tompson 等学者[13]首次将卷积神经网络应用到手势估计中，利用卷积神经网络来提取手部图像特征信息，并为手部关键点生成 2D 热图，然后利用逆运动学原理由热图提取特征，再根据目标函数最小化来估计 3D 手部姿态。这也启发了很多人使用卷积神经网络以及热图进行手部姿态估计。Sinha 等学者[14]利用卷积神经网络获取图像特征的方法，再结合深度数据进行最近邻特征匹配补全手势估计的参数。由于手势估计的复杂性，从图像中估计的关节与真实关节可能出现偏差。

针对上述情况，Ge 等学者[15]先提出了一种新的基于深度图的多视角获取手部关节点后进行回归融合，估算出手势坐标。此后 Ge 等学者[16]根据 Qi 等学者[17]的启发将 PointNet++应用解决三维手势估计问题，将手势深度图 3D 点云进行采样和归一化输入到 PointNet 网络中，进行点云特征提取，同时该方法还设计了一个指尖矫正网络进行指尖位置的优化。随后，Ge 等学者[18]又进一步改变了网络结构，采用编解码器两层架构代替分层架构的采样，对 3D 关节位置进行预测，提高手势估计的精度。在此之前，大多数手部估计方法止步于三维手部关键点的回归，并不能精准地反映手部形态，而 AR/VR 领域需要更加逼真的手部模型。同时，图神经网络能够解决复杂的结构关系，学者将图神经网络引入手势中。因此，Ge 等学者[19]提出一个全新的端到端训练的图卷积神经网络，将 2D 热图等潜在特征变量通过该网络生成了密级手部网格，根据网格坐标最终得到三维关节坐标，原理如图 5 所示。Fang 等学者[20]也提出了基于图卷积网络的联合图推理来估算关节的复杂关系，同时通过增强像素的能力，估算出每个像素的偏移量，再对所有的像素进行加权计算，进而估计出手部信息。

2.2 混合法

生成法又称为基于模型的方法，主要是基于固定的手势模型进行姿态估计识别，需要根据运动学原理事先创建满足手部形态学约束的模型，再进行匹配。主要流程如下：首先需要根据输入图像匹配适合的手部模型，然后进行模型参数初始化，并找到一个实际模型与输入模型之间的损失函数，通过不断迭代最小化损失函数得到最优手势模型。生成法的主要优化方法体现在目标函数最小化方法以及使用先验手势来匹配数据的方法，在本文不进行详细介绍。为了最优化地使用生成法与判别法，有学者提出了混合法，可以使用判别法对姿态进行先验，引导对生成模型的优化，然后使用生成法细化手型与位置，降低跟踪的误差，提高复杂场景环境下跟踪估计的鲁棒性。

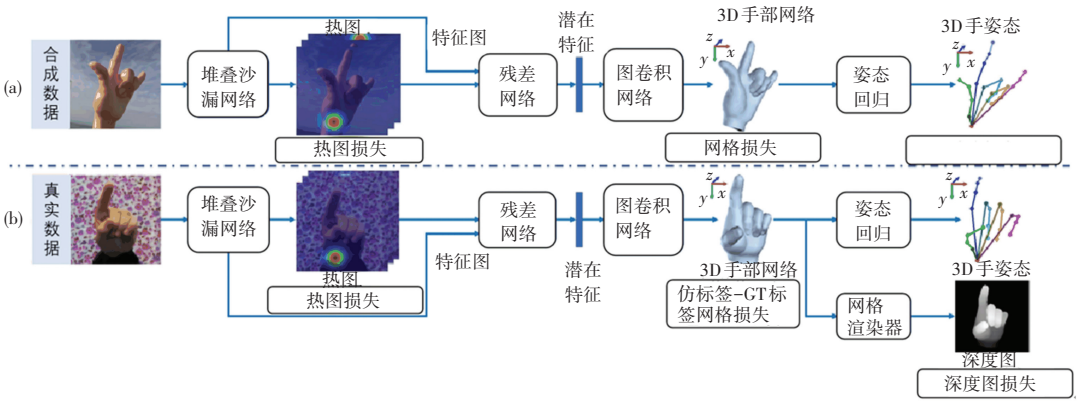

图 5　Ge 等学者[19]提出的网络原理图

Fig. 5　Schematic diagram of the network proposed by Ge et al[19]

Ye 等学者[21]提出基于层次的混合手势估计方法，通过变换输入空间与输出空间的方式，将多阶段与多层回归集成到 CNN 中，在多层级之间，通过粒子群算法把运动学约束施加到 CNN 中，该方法可以减少关节与视角的变化，纠正手势估计的结果。

Mueller 等学者[22]先利用卷积神经网络定位手关节，再使用深度值计算得出手的三维信息。Zhang 等学者[23]先对深度图中的手掌进行分割，并通过预训练的 LSTM 预测当前的手势，最后重建对象模型。

3　数据集与评价指标

3.1　数据集

大规模精准标注的数据集是手势估计的基础，而早期由于缺少专业相机方阵，数据集较小。随机光学组件相关硬件以及计算机软件的发展，使得手势估计数据集已经非常丰富，不仅有手动标注数据、自动标注数据、半自动标注数据，还有全自动合成数据[24]，无论在数据质量、还是数据规模上已经有质的飞越。

手动标记数据有 Dexter-1、MSRA14 等，由于手工标注数据是一件繁琐的事，因此该类数据集规模相对较小，不适合用于大规模数据驱动的手势估计。半自动标注的手势数据有 ICVL、MSRA15、NYU 等，半自动标注方法一般先估算出三维手部关节点，再使用人工标注方法进行修正或者于初始先手动标注出二维手部关节点，再使用算法预测出三维手部关节点，即使使用半自动标注，收集以及标注大数据集的手势数据也是一个繁琐复杂的大工程。为了获得更高质量、更大规模的数据集，出现了全自动以及合成数据集方法。全自动标注数据有 HandNet、BigHand2.2M 等，全自动标注数据先让受试者带上数据手套，在采集图像时进行手部关节数据标注，相较于半自动标注来说自动标注效率大大提高，适合创建大型手势标注数据集。合成数据有 MSRC、RHD 等，合成数据使用软件先基于手势模型生成不同姿态的仿真图像数据，再自动标记三维关节信息。合成数据标记效率高，可以创建大规模的数据集，但合成数据很难对真实图像的丰富纹理特征进行建模，而且因为反关节等各种原因导致数据特征丢失，同时受限于手部的多自由度以及手部肤色，因此就目前来说，合成数据质量相对不高，但随着计算机相关学科的发展，合成数据必将是手势标注数据的发展方向。

表 1 列出了手势估计公共数据集，随着时间的进行，数据量整体呈现上升趋势，从中挑选一个合成数据集、一个超大型数据集以及一个中文手语数据集进行介绍。

（1）RHD（Rendered Hand Pose）。是一个 41 258个训练集以及 2 728 个测试集的手势估计的图像数据集，是由弗莱堡大学在 2017 年发布的合成渲染数据集，每个样本共有深度图、RGB 图、分割图，图像像素为 320 × 320。每只手都有 21 个关键点的精确二维以及三维注释。

（2）FreiHand。是一个包含 32 个人进行的手部动作采集，共有 32 560 个训练样本以及 3 960 个测试样本图像数据集。是由弗莱堡大学与 Adobe 研究院于 2019 年发布的，可用于图像检测、分类任务。

（3）InterHand2.6M。是第一个具有准确 GT 3D 双手交互的大规模手部实拍数据集。由 Facebook Reality Lab 于 2020 年发布，包括 260 万张手势图像。可为学者提供了一个双手交互的手势估计数据集。

表 1 三维手势估计常用数据集

Tab. 1 Common data set of 3D gesture estimation

数据集	时间	图像数量	标记方式	尺寸
STB[25]	2015	36 000	手动	640×480
MSRA14[26]	2014	2 400	手动	320×240
Dexter1[27]	2013	2 137	手动	320×240
InterHand2.6M[28]	2020	260 W	半自动	512×334
FreiHand[29]	2017	133 000	半自动	224×224
MSRA15[30]	2015	76 375	半自动	640×480
NYU[13]	2014	81 009	半自动	640×480
ICVL[31]	2014	17 604	半自动	320×240
HandNet[32]	2015	212 928	自动	320×240
BigHand2.2M[33]	2017	2.2 M	自动	640×480
MSRC[34]	2015	102 000	合成	512×424
RHD[35]	2017	43 700	合成	320×320

3.2 评价指标

手势评价的标准是指相对于标注的手势点相差多少。常见的评价指标可分述如下。

(1)平均关节位置误差(Mean PerJoint Position Error, *MPJPE*)[36], 定义为预测关节点位置与真实三维关节点位置的平均欧几里得距离,单位为 mm。指标值越小、姿态估计算法越好, 计算公式如下:

$$MPJPE_j = \frac{\sum_i (\| p_{ij} - p_{ij}^{gt} \|)}{N} \tag{1}$$

其中, N 表示手指节点数; p_{ij} 表示预测点; p_{ij}^{gt} 表示真实标注点。

(2)端点误差(End Point Error, *EPE*)[37]。定义为手部跟关节对齐后预测的三维手部坐标与真实坐标之间的平均欧式距离,单位为 mm。计算公式如下:

$$EPE = \frac{\sum_{m=1}^{S} (\sum_{k=1}^{i} \left\| \frac{(y_{mk} - y_{m0}) - \hat{y}_{mk} - \hat{y}_{m0}}{\max(w,h)} \right\|)}{i \times S} \tag{2}$$

其中, S 为样本数; i 为关节点数; y 表示真实值; $\hat{y}$ 表示预测值。

(3)正确关键点百分比(Percentage of Correct KeyPoints, *PCK*)[38] 表示手势估计结果预测值与真实值相差的欧氏距离在一定可接受范围内,则认定为预测准确。J_k 计算公式如下:

$$PCK_i^k = \frac{\sum_p \delta\left(\frac{d_{pi}}{d_p^{dcf}} \leqslant T_k\right)}{\sum_p 1} \tag{3}$$

其中, T_k 表示阈值。

(4)工作特征曲线下面积(Area Under Curve, *AUC*)[39]。在手势估计中,*AUC* 被定义为 *PCK* 曲线与坐标轴围成的面积,相同标准下 *AUC* 值越大表示估计误差越小,精度越高。

不同算法在 RHD 以及 STB 公开数据集上执行精度对比见表 2。

表 2 不同算法的精度比较

Tab. 2 Precision comparison of different algorithms

方法	*AUC* (STB)	*AUC* (RHD)
Krejov 等[40]	0.991	0.849
Yang 等[41]	0.996	0.901
Ge 等[19]	0.998	0.920
GU 等[42]	0.996	0.887
Mceu 等[43]	0.965	0.560
Zhou 等[44]	0.991	0.893
Chen 等[45]	0.990	0.939

4 问题与挑战

当前已经有较多的学者参与研究三维手势估计,基于单目 RGB、双目、RGB-D 的估计在特定场景设备下已经取得了较大进步,但是在特殊环境进行复杂操作时仍然有较多的问题亟待解决,例如:环境背景与手掌肤色贴合、光照变化较大、进行复杂的自遮挡动作等[46]。

4.1 复杂场景环境

为了精准分割出手势图像,大部分手势估计方法均在背景单一、且单手条件下进行,而正常环境下可能无法控制在环境光照变化较强的场景或者与手肤色相近的背景或者反光面、玻璃等背景下的多手协作。因为,高光照在这种复杂的背景环境中无疑加大了手势检测、分割的难度。例如:强光照射手部或阴影投射手部均使手与背景不明显。如何提高手势估计在复杂场景背景下的手势检测与分割的精准性,进而提高复杂场景的手势交互能力,将会是未来的一个研究方向。

4.2 高自由度

人手有 26 个自由度,可以实现 300°/s 旋转以及5 m/s的快速运动,因此十分灵活,手势估计姿态的复杂度随着自由度以及运动速度的增加而呈指数的增长。目前仍存在较多精度较低、无法贴合手部结构的运动模型。如何在高自由度的快速运动的手部图像序列中进行精准识别高维时序特征,快速预测手部关节值仍然是一个热点问题。

4.3　自遮挡

因为手部的高自由度导致手部具有多样性以及多异性。人类很容易实现的自握拳、自握手等无疑会出现手部自遮挡、自碰撞。而且因为肤色、年龄等差异较大，加上自遮挡问题，可能使得手部在图像中所占面积较小，进而丢失较多手部细节信息，导致手势估计不准确或者完全失效。

4.4　实时性与准确性

当前较多研究是在实验室环境中使用高性能计算机进行检测、分割，其运行速率可达 90 FPS 以上，而在手机或者 AR 眼镜上，加上复杂的环境等因素，其处理速度可能达不到 10 FPS，AR/VR 应用的理想运行速率不低于 60 FPS。因此，在复杂的环境下，需要实现准确性与实时性，仍然有较多问题需要解决。

5　展　望

基于深度学习的三维手势估计方法不断进行优化，极大地提升了手势估计的效果，基于上文提出的问题，研究者可以从以下几个方面进行优化。

5.1　利用时序信息

基于时间序列的手势估计可以利用双向长短时记忆网络模型获取前后帧之间的时序特征，挖掘出更加丰富的特征信息，进而辅助预测出后续手掌位置、甚至手势关键节点信息，解决自遮挡等复杂环境背景下手势识别的准确性以及手势估计的速度问题。

5.2　优化网络模型

深度学习的手势估计中，网络模型是一个重要的主题。如何优化出轻量级的网络模型解决复杂的场景下手势检测与分割以及特征提取等手势估计的准确性问题，进而提高网络的运行速度，是助力手势估计研究的一个重要学术方向。

5.3　利用混合法

判别法对遮挡等有较强的鲁棒性问题可以快速从错误中恢复，而且其运行速度较快，但是却无法利用时序帧，导致手势估计容易出现跟踪丢失现象，而生成法可以利用时序帧，使用拟合模型处理高维数据和复杂环境下的手势估计。如何平衡使用判别法与混合法，充分利用二者的优势，可加快手势估计跟踪的性能。

6　结束语

本文对基于深度学习的手势估计算法以及数据集和评价指标进行了回顾，探讨了手势估计目前所面临的挑战以及未来的研究方向。手势交互是最重要的人机交互之一，应用在 AR/VR、手语识别、远程操控等方面，虽然不少学者在手势估计方面的研究已经取得了一定成果，但是距离实际应用还有较长的路要走。因此，也希望相关研究学者继续进行复杂场景的手势研究，让手势估计早日在中低端设备上落地应用。

参考文献

[1] 解迎刚，王全. 基于视觉的动态手势识别研究综述[J]. 计算机工程与应用，2021，57(22)：68-77.

[2] KRIZHEVSKY A，SUTSKEVER I，HINTON G E. Imagenet classification with deep convolutional neural networks [J]. Communications of the ACM，2017，60(6)：84-90.

[3] 易靖国，程江华，库锡树. 视觉手势识别综述[J]. 计算机科学，2016，43(S1)：103-108.

[4] JIANG Linjun，XIA Hailun，GUO Caili. A model-based system for real-time articulated hand tracking using a simple data glove and a depth camera[J]. Sensors，2019，19(21)：4680-1788.

[5] QIAN Jing，MA Jiajun，LI Xiangyu，et al. Portalble：Intuitive free-hand manipulation in unbounded smart phone-based augmented reality[C]//Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology. Montreal，Canada：ACM，2019：133-145.

[6] 方林普. 基于视觉的手势交互关键技术研究[D]. 广州：华南理工大学，2021.

[7] 陈红梅，赖重远，张洋，等. 基于深度数据的手势识别研究进展[J]. 江汉大学学报(自然科学版)，2018，46(2)：101-108.

[8] DOOSTI B. Hand pose estimation：A survey[J]. arXiv preprint arXiv：1903.01013，2019.

[9] 武国梁. 基于深度学习的手势估计研究[D]. 长春：中国科学院大学(中国科学院长春光学精密机械与物理研究所)，2021.

[10] 王健. 基于深度学习的手势识别算法研究[D]. 长春：长春理工大学，2021.

[11] EROL A，BEBIS G，NICOLESCU M，et al. Vision-based hand pose estimation：A review [J]. Computer Vision and Image Understanding，2007，108(1/2)：52-73.

[12] 张继凯，李琦，王月明，等. 基于单目 RGB 图像的三维手势跟踪算法综述[J]. 计算机科学，2022，49(4)：174-187.

[13] TOMPSON J，STEIN M，LECUN Y，et al. Real-time continuous pose recovery of human hands using convolutional networks[J]. ACM Transactions on Graphics，2014，33(5)：1-10.

[14] SINHA A，CHOI C，RAMANI K. DeepHand：Robust Hand Pose Estimation by Completing a Matrix Imputed with Deep Features [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas，NV，USA：IEEE，2016：4150-4158.

[15] GE Liuhao，LIANG Hui，YUAN Junsong，et al. Robust 3d hand pose estimation in single depth images：from single-view cnn to multi-view CNNs[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，Las Vegas，USA：IEEE，2016：3593-3601.

[16] GE Liuhao，CAI Yujun，WENG Junwu，et al. Hand pointnet：3d hand pose estimation using point sets[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City：IEEE，2018：8417-84.

[17] QI C R，YI Li，SU Hao，et al. Pointnet++：Deep hierarchical

feature learning on point sets in a metric space[C]//Advances in Neural Information Processing Systems. Long Beach, California, USA: NIPS Foundation,2017: 5099-5108.
[18] GE Liuhao, REN Zhou, YUAN Junsong. Point - to - point regression pointnet for 3d hand pose estimation[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany:dblp, 2018: 475-491.
[19] GE Liuhao, REN Zhou,LI Yuncheng,et,al. 3D Hand shape and pose estimation from a single RGB image[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Long Beach:IEEE, 2019:10833-10842.
[20] FANG Linpu, LIU Xingyan, LIU Li, et al. JGR-P2O: Joint graph reasoning based pixel-to-offset prediction network for 3D hand pose estimation from a single depth image[C]//European Conference on Computer Vision. Cham:Springer, 2020: 120-137.
[21] YE Qi,YUAN Shanxin, KIM T K. Spatial attention deep net with partial pso for hierarchical hybrid hand pose estimation[C]// European Conference on Computer Vision. Cham: Springer, 2016:346-361.
[22] MUELLER F, MEHTA D, SOTNYCHENKO O.et al. Real-time hand tracking under occlusion from an egocentric RGB-D sensor [C]//2017 IEEE International Conference on Computer Vision. Venice, Italy:IEEE, 2017: 1163-1172.
[23] ZHANG Hao, BO Zihao, YONG Junhui, et al. Interaction Fusion: Real-time reconstruction of hand poses and deformable objects in hand-object interactions[J]. ACM Transactions on Graphics (TOG),2019,38(4):1-11.
[24] 王丽萍,汪成,邱飞岳,等. 深度图像中的 3D 手势姿态估计方法综述[J]. 小型微型计算机系统,2021,42(6):1227-1235.
[25] WHEATLAND N, WANG Yingying, SONG Huaguang, et al. State of the art in hand and finger modeling and animation[J]. Computer Graphics Forum, 2015, 34(2): 735-760.
[26] QIAN Chen,SUN Xiao, WEI Yichen, et al. Realtime and robust hand tracking from depth[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA:IEEE,2014: 1106-1113.
[27] SRIDHAR S, OULASVIRTA A, THEOBALT C. Interactive markerless articulated hand motion tracking using RGB and depth data[C]//Proceedings of the IEEE International Conference on Computer Vision. Australia:IEEE,2013: 2456-2463.
[28] MOON G, YU S I, WEN He, et al. InterHand2. 6M: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image]J]. arXiv preprint arXiv:2008.09309,2020.
[29] ZHANG Jiawei, JIAO Jianbo, CHEN Mingliang, et al. A hand pose tracking benchmark from stereo matching[C]//2017 IEEE International Conference on Image Processing (ICIP). Beijing, China:IEEE, 2017:982-986.
[30] SUN Xiao,WEI Yichen, LIANG Shuang, et al. Cascaded hand pose regression[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Xi'an, China:IEEE, 2015: 824-832.
[31] TOMPSON J, STEIN M, LECUN Y, et al. Real-time continuous pose recovery of human hands using convolutional networks[J]. ACM Transactions on Graphics,2014,33(5):1-10.
[32] TANG Danhang, JIN C H, TEJANI A,et al. Latent regression forest:Structured estimation of 3D articulated hand posture[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus:IEEE,2014: 3786-3793.
[33] WETZLER A, SLOSSBERG R, KIMMEL R. Rule of thumb: Deep derotation for improved fingertip detection[EB/OL]. [2015]. https://arxiv.org/abs/1507.05726.
[34] YUAN Shanxin, YE Qi, STENGER B, et al. Bighand2. 2m benchmark: Hand pose dataset and state of the art analysis[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Venice, Italy:IEEE,2017: 4866-4874.
[35] SHARP T, KESKIN C, ROBERTSON D, et al. Accurate, robust,and flexible real-time hand tracking[C]//Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. Seoul:ACM,2015:3633-3642.
[36] ZIMMERMANN C,BROX T. Learninglo estimate 3d hand pose from single rgb images[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy:IEEE, 2017:4903-4911.
[37] LOBO J M, JIMENEZ-VALVERDE A, REAL R. AUC: A misleading measure of the performance of predictive distribution models[J]. Global ecology and Biogeography, 2008, 17(2): 145-151.
[38] MCKEE I W, WILLIAMSON P C, LAM E W, et al. The accuracy of 4 panoramic units in the projection of mesiodistal tooth angulations[J]. American journal of orthodontics and dentofacial orthopedics, 2002, 121(2): 166-175.
[39] VINT P F, HINRICHS R N. Endpoint error in smoothing and differentiating raw kinematic data: An evaluation of four popular methods[J]. Journal of Biomechanics, 1996, 29(12): 1637-1642.
[40] KREJOV P, BOWDEN R. Multi-touchless: Real-time fingertip detection and tracking using geodesic maxima[C]// 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). Shanghai, China:IEEE,2013:1-7.
[41] YANG Linlin, YAO A. Disentangling latent hands for image synthesis and pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach :IEEE,2019:9877-9886.
[42] YANG Linlin, LI Shile, Lee D,et al. Aligning latent spaces for 3d hand pose estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul:IEEE,2019: 2335-2343.
[43] GU Jiajun,WANG Zhiyong, OUYANG Wanli,et al. 3d hand pose estimation with disentangled crossmodal latent space[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Snowmass Village:dblp, 2020: 391-400.
[44] MCEU E R F, BERNARD F, SOTNYCHENKO O, et al. GANerated hands for real-time 3d hand tracking from monocular RGB[J]. arXiv preprint arXiv:1712.01057,2017.
[45] ZHOU Yuxiao, HABERNANN M, Xu Weipeng, et al. Monocular real-time hand shape and motion capture using multi-modal data[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle :IEEE, 2020: 5346-5355.
[46] CHEN Liangjian,L1N S Y,XIE Yusheng,et al. DGGAN:Depth-image guided generative adversarial networks for disentangling rgb and depth images in 3d hand pose estimation[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Seattle : IEEE,2020:411-419.
[47] 梁晓辉. 手部姿态估计方法综述[J]. 山西大学学报(自然科学版),2022,45(3):631-640.