基于局部深度一致性的自监督手部姿态估计

王敬宇[1],黄伟亭[1],刘 聪[2],戚 琦[1],孙海峰[1],廖建新[1]
(1. 北京邮电大学网络与交换国家重点实验室,北京 100876;2. 中国移动通信有限公司研究院,北京 100053)

摘 要: 基于深度图的3D手部姿态估计通常需要大量人工标注数据以达到高精确度和鲁棒性,然而关节点标注过程冗杂且存在一定误差. 现有研究工作使用自监督方法解决对标注数据的依赖,通过在虚拟数据集上预训练网络,并在无标注的真实数据集上进行模型拟合,实现3D姿态估计. 自监督方法的关键在于设计模型拟合的能量函数以减小模型在真实数据集上的精度下降程度. 为了减小模型拟合难度,本文提出局部深度一致性损失,依据初始姿态估计结果,提取输入与输出深度图的局部表征,将深度图显式地解耦为以关节点为中心的不同区域. 通过有针对性地对不同关节点进行局部优化,减少虚拟与真实深度图之间的固有领域误差对网络学习的影响,增加训练的稳定性. 本文方法在NYU数据集上相比基础方法平均关节点误差提升了21.9%.

关键词: 自监督;手部姿态估计;局部一致性;深度图;深度学习
基金项目: 国家重点研发计划(No.2020YFB1807800);国家自然科学基金(No.62071067, No.62001054, No.61771068);教育部-中国移动科研基金(No.MCM20200202, No.MCM20180101);博士后创新人才支持计划(No.BX20200067);中国博士后科学基金资助(No.2021M690469)
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2023)06-1644-10
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20210648

Self-Supervised Hand Pose Estimation with Regional Depth Correspondence

WANG Jing-yu[1], HUANG Wei-ting[1], LIU Cong[2], QI Qi[1], SUN Hai-feng[1], LIAO Jian-xin[1]
(1. *State key laboratory of Networking and Switching Techonology, Beijing University of Posts and Telecommunications, Beijing* 100876, *China*;
2. *China Mobile Group Design Institute Co., Ltd., Beijing* 100053, *China*)

Abstract: Depth-based 3D hand pose estimation requires manually labelled data to achieve high accuracy and robustness. However, the labeling process is laborsome and bares inevitable biases. Researchers solve this problem by using self-supervised methods. They pretrain model on synthetic dataset then finetune on unlabelled real dataset through model fitting. The biggest challenge is the design of model fitting term in fintuning stage to prevent severe accuracy drop. We proposed the regional depth correspondence loss which utilized initial pose estimation results to extract regional representation of input and output depth maps and transparently divided them into different regions. This allows network to finetune regions around joints without being affected by overall domain gaps between synthetic and real depth images. The proposed method outperforms baseline method by 21.9% on NYU hand pose dataset.

Key words: self-supervised; hand pose estimation; regional consistency; depth images; deep learning
Foundation Item(s): National Key Research and Development Program of China (No.2020YFB1807800); National Natural Science Foundation of China (No.62071067, No.62001054, No.61771068); Ministry of Education-China Mobile Research Foundation (No. MCM20200202, No. MCM20180101); Postdoctoral Innovative Talent Support Program (No.BX20200067); China Postdoctoral Science Foundation (No.2021M690469)

1 引言

手是人类与外部世界交互的主要媒介之一[1],手部姿态估计是人机交互[2~4]、虚拟现实[5]、机器人[6]等领域的重要研究热点. 深度学习与商用摄像头的出现大大

收稿日期:2021-05-20;修回日期:2021-08-15;责任编辑:王天慧

促进了基于深度图的3D手部姿态估计方法[7~11]的发展. 这类任务的目标是输入一张包含手的深度图,输出手上主要关节点的3D坐标. 其中深度图上每个点表示空间中该点到摄像头的距离,单位为米或毫米. 然而,这类数据驱动(Data driven)方法要求对关节点坐标及相关的特征图进行监督训练,其效果与标注数据集大小、数据多样性和标注的准确度成正相关[9,12],即使是在深度图这种深度歧义较少的表征上,标注手部关节点的3D信息仍然需要耗费巨大的人力. 开源3D手部姿态数据集通过多视角跟踪[13]或者6DoF传感器[12]等方法获取关节点坐标信息,需要细致地标定、对齐传感器及摄像头并手动校正偏差较大的标注.

由于真实数据集标注困难且存在随机误差,许多领域开始探索自监督方法,通过在大量虚拟数据集上对网络进行预训练,使网络获得一定的特征提取能力,而后在无标注的真实数据集上拟合网络,这样能够在保持较少精度下降程度的前提下摆脱网络对真实标注数据的依赖. 对于基于深度图的自监督手部姿态估计任务,得益于3D渲染技术的发展,仅有的三个相关工作[14~16]均是在输出维度上,对网络输出的深度图与输入深度图进行对齐. 如图1所示,输出深度图(虚拟)平滑无噪声,而输入深度图(真实)带有随机噪声和深度值缺失,由于数据处理方式不同,还会出现手部不完全匹配的情况,例如真实深度图带有手腕和手臂,而虚拟深度图只有手掌和手指部分. 不同于有监督方法中使用关节点到关节点的强监督信息解决关节点回归问题,自监督方法中只能使用输入与输出深度图的弱监督信息,同时解决关节点估计误差和真实深度图与虚拟深度图的领域误差(Domain gap)问题,在这种情况下,Dibra等[15]和Wan等[16]直接对整体输入输出深度图计算损失,将不同关节点的误差与领域误差耦合到一起,没有充分利用手的空间结构信息对关节点进行针对性优化,增加网络学习难度,尤其在遮挡情况下,网络的拟合能力减弱,网络训练不稳定. Wan等[14]通过学习深度图到手模型的稠密对应图(Dense correspondence map),为输入输出深度图加入一对一的对应关系,然而这种方法强依赖稠密对应图估计的准确率,在上述提到的真实与虚拟深度图领域误差的情况下,精确估计上千个点的坐标对网络的要求较高,给后续优化带来难度. 两个相关工作[14,16]都使用多视角信息给网络加入更多监督信息,能有效提高网络精度,但多个摄像头的使用与标定削弱了方法在实际场景中的应用范围.

为了解决输入输出深度图直接对齐带来的网络优化难度增加的问题,本文提出局部深度一致性损失,如图1所示,本方法依据初始姿态估计结果,将输入、输出深度图解耦合为以关节点为中心的局部区域(为了可视化效果呈现了完整的深度图,实际计算损失的部分为深度图中高亮的圆形区域),计算局部一致性损失,这样有针对性地进行局部优化,可以减小输入的真实深度图与输出的虚拟深度图之间固有的领域误差对网络训练的影响,使网络训练更稳定,即使在遮挡情况下也能进行合理的估计. 局部深度的划分在一定程度上依赖初始姿态估计准确率,但在实验部分(3.2节)证明所提方法在初始估计略有偏差的情况下仍然具备一定的校正能力. 本文在开源NYU[2]手部姿态估计数据集上进行了充分实验,平均关节点误差相比基础方法提升了21.9%,证明了所提方法的有效性. 另外还设计实验,考虑数据集与手模型因关节点定义不同而造成的偏差,在去掉偏差后所提方法的真实效果相比基础方法提升了22.2%.

图1　局部深度表征

2　相关研究

随着神经网络的发展与商用深度摄像头的出现,基于深度图的3D手部姿态估计的精度和速度获得了大幅度提升,这些方法可以分为三类:生成方法、判别方法和混合方法.

生成方法通过迭代地优化目标函数,达到拟合手模型和输入图片的效果. Melax S等[17]将问题制定为受约束的刚体问题并使用迭代最近点算法(Iterative Closest Point,ICP)解决;Tompson J等[13]使用非梯度离线粒子群算法(Particle Swarm Optimization,PSO)拟合模型得到NYU数据集的关节点标注. 生成方法不需要标注数据,对未知数据泛化效果较好,但是对初始化敏感,用在手部追踪领域误差随时间累积,当误差达到一定程度需要重新初始化.

判别方法通过机器学习、深度学习方法直接预测手部相关参数,如关节点坐标、转角等. Wan等[7]设计了关节点偏移向量场,充分利用深度图的特性,编码手的空间结构信息;Huang等[8]在此基础上提出自适应加权模块对特征图上的信息进行聚合,使网络可以端到端训练;为了充分利用深度图2.5D的特性,Chen等[9]将深度图转化为点云,Moon等[11]将深度图转化为体元素并使用3DCNN进行处理.

混合方法使用判别方法进行初始化,而后使用生成方法进行模型拟合、优化. Sinha等[18]使用CNN降低

深度图的维度，而后通过矩阵补全方法（Matrix completion approach）优化最终姿态；Oberweger等[19]使用深度生成网络虚拟深度图像，再使用单独的优化网络迭代地校正手部姿态.

有监督姿态估计中判别方法表现较好[20]，但是需要大量准确的标注数据. 基于引言所述获取真实3D标注数据的难处，领域内开始探索不依赖人工标注数据的自监督手部姿态估计方法. Dibra等[15]首次提出利用虚拟数据集训练模型，而后在无标签的真实数据上进行拟合的自监督训练流程，这种方法既能直接利用现有判别式模型，保证网络推理速度，也能降低模型对数据的依赖. 在此训练流程的基础上，Wan等[16]提出利用多视角信息增强网络拟合能力. 这两种方法在真实数据拟合阶段均直接对齐完整的输入、输出深度图，网络学习受虚拟数据与真实数据之间的领域误差所影响，增加网络训练难度与稳定性. Wan等[14]进一步加入手模型与深度图之间的稠密对应关系，实验证明这种点对点的自监督关系给网络带来了巨大的提升. 但是构建手模型与深度图之间上千个点的对应关系提升了网络的复杂度.

基于现有数据驱动方法的有效性，本文方法使用Huang等[8]的工作实现关节点坐标估计任务（3.2节），为了实现网络在真实数据上进行自监督拟合的任务，提出手模型参数回归网络（3.3节）估计手模型参数得到手模型网格化蒙皮，通过对手模型进行3D渲染得到输出深度图，通过利用前一阶段获得的关节点坐标对输入、输出深度图进行解耦合，将深度图划分为以关节点为中心的不同区域，计算局部深度一致性损失，通过对关节点局部区域的针对性优化，不断调整关节点坐标，在保证网络稳定训练的同时达到良好的关节点估计效果.

3 研究框架

本节先讨论所提出方法的整体网络结构设计（3.1节），包括特征提取网络、3D关节点坐标估计（3.2节）与手模型参数估计分支（3.3节），随后介绍以及网络训练时使用的局部深度一致性损失（3.4节）和碰撞损失（3.5节），最后介绍网络整体训练流程（3.6节）.

3.1 整体结构

本文所提方法的整体网络结构如图2所示，图中所使用的符号及含义如表1所示. 整体网络由特征提取网络、3D关节点坐标估计分支与手模型参数估计分支组成. 特征提取网络采用编码器-解码器（Encoder-decoder）结构，其中编码器部分使用ResNet-50[21]，考虑到手的局部特性，需要较小的卷积核尺寸提取局部特征，于是将预处理部分的卷积核尺寸由7改为5[8,10]，并去掉网络最后的均值池化和全连接层；解码器部分由3层转置卷积层组成，将特征上采样到原图像的1/2. 输入深度图D_i经过该特征提取网络得到关节点3D偏移

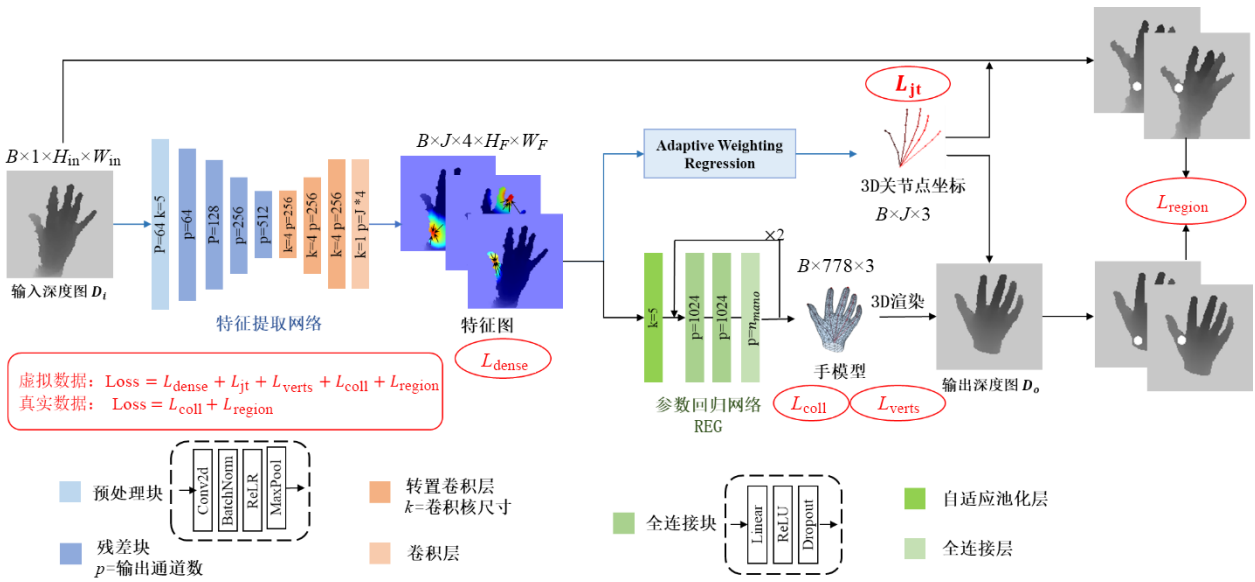


图2 整体网络结构

表1 图2中所使用符号及其含义

符号	含义
B	Batch size,批尺寸
$\boldsymbol{D}_i$	输入深度图
$\boldsymbol{D}_o$	输出深度图
(H_{in}, W_{in})	输入图片分辨率,本文为(128,128)
(H_F, W_F)	特征图分辨率,本文为(64,64)
K	关节点数量,本文为21

向量特征图[7,8].

特征图后接两个分支,其中3D关节点估计分支(3.2节),采用Adaptive Weighting Regression(AWR)模块[8]对特征图进行处理,得到 $J=21$ 个关节点的3D坐标;手模型参数估计分支(3.3节),迭代地回归手模型参数,经过前向骨骼传递(Forward kinematics)算法处理后得到包含778个顶点的手部网格化蒙皮模型,经过3D渲染后可输出虚拟深度图 D_o.

注意在训练完成后,推理阶段仅保留特征提取网络和3D关节点估计分支(图2中蓝色箭头连接部分).

网络训练过程中使用到5个损失函数,其中 L_{dense}, L_{jt}, L_{verts} 分别是3D偏移向量特征图损失和3D关节点坐标损失,需要3D关节点和手模型顶点标注,仅在虚拟数据集上使用;L_{region}, L_{coll} 分别是本文所提的局部深度一致性损失和碰撞损失,无需标注信息,在虚拟数据和真实数据的训练过程中均使用.

3.2　3D关节点坐标估计(AWR)

3D关节点坐标估计是本文方法的主要任务. 考虑到深度图的空间特性和手的结构性,本文使用3D偏移向量特征图[7]作为3D关节点的稠密表征(Dense representation). 具体地,3D偏移向量特征图的维度如图2所示,其上每个点表示空间中该点到某个特定关节点的偏移量(Offset),用3维度的单位方向向量 $\boldsymbol{V}\left(p_i, p_j\right)$ 和1维度的"亲近"热图[8](Closeness heatmap) $\boldsymbol{S}\left(p_i, p_j\right)$ 表示. 表达式如式(1)所示,其中 $1_{\text{Hand}}\left(p_i\right)$ 是指示函数,当点 p_i 在手的区域内,其值为1,否则为0;$p_j(j\in\{1,2,\cdots,K\})$ 表示3D关节点坐标;k 表示关节点周围偏移量特征的半径,超出该范围的点不考虑其对计算关节点坐标的贡献. 在本文中,取 $k=1$ [8].

$$\boldsymbol{S}\left(p_i, p_j\right)=\begin{cases}1_{\text{Hand}}\left(\boldsymbol{p}_i\right)\cdot\dfrac{k-\left|p_i-p_j\right|}{k}, & \left|p_i-p_j\right|<k\\ 0, & \text{otherwise}\end{cases}$$

$$\boldsymbol{V}\left(p_i, p_j\right)=\begin{cases}1_{\text{Hand}}\left(p_i\right)\cdot\dfrac{p_i-p_j}{k}, & \left|p_i-p_j\right|<k\\ 0, & \text{otherwise}\end{cases}\tag{1}$$

这种坐标表示方式采用了合理的3D距离表征,特征图上每个点不仅能表示该点对某个关节点的"亲近"程度,也能表征关节点的相对位置,显式地编码了手的结构.

基于3D偏移向量的物理意义,特征图上每个点的坐标加该点到某个关节点的偏移向量,即能获得该关节点的坐标,因此自适应加权模块(Adaptive Weighting Regression, AWR)模块利用"亲近"热图对特征图上每个点计算出的关节点坐标进行加权聚合,即可以可导方式获得关节点坐标,如式(2)所示.

$$p_j=\boldsymbol{S}\left(p_i, p_j\right)\cdot\left[p_i+\boldsymbol{V}\left(p_i,p_j\right)\left(k-k\cdot\boldsymbol{S}\left(p_i, p_j\right)\right)\right]\tag{2}$$

该分支在虚拟数据上训练时对特征图和3D关节点坐标进行监督,分别表示为 L_{dense}, L_{jt},如式(3)所示,带 * 项表示标注信息(Groundtruth).

$$\begin{aligned}L_{\text{dense}}=&\frac{1}{K}\left(\sum_{i=1}^{K}\text{smooth}_{L1}\left(\boldsymbol{S}^*\left(p_i,p_j\right),\boldsymbol{S}\left(p_i, p_j\right)\right)+\right.\\&\left.\text{smooth}_{L1}\left(\boldsymbol{V}^*\left(p_i,p_j\right),\boldsymbol{V}\left(p_i, p_j\right)\right)\right)\\L_{\text{jt}}=&\frac{1}{K}\sum_{i=1}^{K}\text{smooth}_{L1}\left(p_j^*, p_j\right)\end{aligned}\tag{3}$$

3.3　手模型参数估计(REG)

该分支的功能是通过网络估计手模型参数以拟合手模型,进而渲染出虚拟深度图,为真实数据提供自监督信息. 本文使用由全连接层组成的回归模块估计手模型参数,然而由特征图到一维手模型主成分参数之间的映射存在较大的语义鸿沟(Semantic gaps). 相关工作[22,23]发现采用由粗糙到细致(Coarse to fine)的迭代结构能不断优化估计结果. 因此本文实现了迭代回归模块,从特征图中拟合得到手模型参数. 通过手模型参数以及手模型内置的形状、姿态系数,经过逆向骨骼算法(Inverse Kinematics, IK)及线性蒙皮方法可以得到图2所示的手的网格化模型[24],结合相机参数和3D渲染技术得到对应的输出深度图,与输入深度图计算损失可以实现对网络的自监督.

本文使用手的表征模型MANO[24]是手部表面蒙皮的低维度参数化模型,该模型是在31位候选人约1 000张展示不同手势的3D扫描图上重建的,因此它能细致地再现手的形状(Shape)和姿态(Pose)变化. 具体来说,MANO手模型输出表面由778个顶点构成的网格化蒙皮 $\boldsymbol{M}\left(\vec{\boldsymbol{\beta}},\vec{\boldsymbol{\theta}}\right)\in\mathbf{R}^{778\times3}$. 如式(4),$\vec{\boldsymbol{\beta}}\in\mathbf{R}^{10}$ 是形状(Shape)参数,表示不同手的形状的主成分系数;$\vec{\boldsymbol{\theta}}\in\mathbf{R}^{51}$ 是姿态(Pose)参数,包含45个关节点转角主成分系数、3个全局旋转系数和3个平移系数;$\boldsymbol{T}(\cdot)$ 是自然状态下(五指伸直微张)的参数化手部模板(Template),具有与手的形状、姿态参数相关的校正系数以保证手模型的真实性;$\boldsymbol{J}(\cdot)$ 用于根据蒙皮顶点坐标计算关节点坐标;$\boldsymbol{\mathcal{W}}$ 是线性蒙皮混合系数矩阵,用于线性混合蒙皮(Linear Blend Skinning, LBS)函数 $\text{W}(\cdot)$ 中计算手的网格化蒙皮. $\text{T}(\cdot)$, $\text{J}(\cdot)$ 都是关于 $\left(\vec{\boldsymbol{\beta}},\vec{\boldsymbol{\theta}}\right)$ 的可导函数,因此MANO对于手模型参数 $\left(\vec{\boldsymbol{\beta}},\vec{\boldsymbol{\theta}}\right)$ 可导,可用于神经网络中进行端到端训练.

$$\boldsymbol{M}\left(\vec{\boldsymbol{\beta}},\vec{\boldsymbol{\theta}}\right)=\text{W}\left(\boldsymbol{T}\left(\vec{\boldsymbol{\beta}},\vec{\boldsymbol{\theta}}\right),\boldsymbol{J}\left(\vec{\boldsymbol{\beta}}\right),\vec{\boldsymbol{\theta}},\boldsymbol{\mathcal{W}}\right)\tag{4}$$

在获得手的网格化蒙皮后,手模型的关节点坐标可以通过蒙皮上顶点坐标的线性插值获得,即

$$p_j=\sum_{i=1}^{778}\boldsymbol{J}_{ij}\boldsymbol{M}_i \tag{5}$$

相比Dibra等[15]和Wan等[14,16]的工作中使用的手模型，MANO具有边缘平滑、细节真实、可表征不同形状、其形状与姿态系数中自带手部先验约束等优点，可有效降低自监督任务的难度. 本文使用MANO手模型生成虚拟数据集，用于网络中以渲染输出深度图. 在网络中，参数回归网络（REG）估计手模型参数 $n_{\text{mano}}=3+45+10+4=62$，分别表示全局旋转、姿态参数、形状参数和相机缩放、平移参数.

3.4 局部深度一致性损失

主流自监督3D手部姿态估计方法[14~16]在无监督真实深度图上优化时，将问题转化为模型优化问题，它们通常直接对整体输入、输出深度图做差值作为网络的能量函数（Energy function），然而如图1所示，真实深度图往往边缘不平滑，在自遮挡严重的情况下存在一定程度的深度值缺失现象[25]（图6），且由于预处理时只对手部进行大致截取，很多情况下输入深度图会包含一部分的手腕及手臂，与只包含手掌和手指的网络渲染的输出深度图存在不可消除的固有误差. 在此情况下直接对深度图整体做差值，将不同关节点的误差与领域误差耦合到一起[25]，增加网络学习难度，无法对关节点进行针对性优化.

本文提出使用关节点坐标估计分支获得的初始姿态估计结果，提取输入、输出深度图的局部信息，显式地将深度图解耦合为 K 个部分，在每个区域中，网络只需要针对当前关节点附近的深度值进行优化，而无需考虑全局深度图之间的固有误差. 如图1所示输入、输出局部深度图，网络估计手腕节点时，只需要考虑手腕附近输入、输出深度值的一致性，无需考虑输入深度图中多余的手臂部分. 由于局部深度的划分与初始姿态估计精度有关，本文的关节点坐标估计分支以3D偏移向量作为手的稠密特征表示，只考虑特定关节点附近的点与关节点之间的距离和方向等局部信息，一定程度上削弱了虚拟数据与真实数据的领域误差.

局部深度一致性损失如式（6）~式（8）所示. 式（6）和式（7）表示局部深度图的构成方式. 首先计算每个关节点 p_j 对应的热图 $H_j(u_i,v_i)$，它是以关节点 p_j 为中心的二维高斯概率分布，以距离度量二维平面上每个像素是关节点的概率. 取值为0~1（式（6））.

取热图中概率大于0.01且在手部范围内的点的位置为掩膜 $\text{Mask}(u_i,v_i)$（式（7））.

利用掩膜分别提取输入 D_{in} 和输出 D_{out} 的局部深度信息，最后计算以 N 个关节点为中心的局部深度的平滑 L_1 损失（Smooth L_1 Loss）（式（8））.

$$\boldsymbol{H}_j(u_i,v_i)=\exp\left(-\frac{\left(u_i-p_{jx}\right)^2+\left(v_i-p_{jy}\right)^2}{2\sigma^2}\right) \tag{6}$$

$$\textbf{Mask}_j(u_i,v_i)=1_{\text{Hand}}(u_i,\ v_i)\cdot 1_{H_j(u_i,v_i)>0.01} \tag{7}$$

$$L_{\text{region}}=\frac{1}{N}\sum_{j=1}^{N}\text{smooth}_{\text{L1}}\left(\textbf{Mask}_j\cdot\left(\boldsymbol{D}_i-\boldsymbol{D}_o\right)\right) \tag{8}$$

3.5 碰撞损失

虽然MANO手模型[24]中的形状和姿态参数中自带手部先验约束，有效地减少了不合理姿态的出现，然而由于网络监督与估计手模型参数任务之间是高度非线性的抽象问题，不可避免地会出现手模型不同部位相互碰撞、穿插的情况，使用碰撞损失[15,16]以规避上述问题.

碰撞损失表示如式（9）所示. 以关节点为圆心 c，每个关节点预设半径 r，当第 i 个关节点位置上的圆球与第 j 个关节点位置上的圆球之间球心的距离小于两个圆球的半径时，视两个关节点相互碰撞.

$$L_{\text{coll}}=\sum_{i,j,i\neq j}\max\left(r_i+r_j-\left|c_i-c_j\right|_2^2,\ 0\right) \tag{9}$$

3.6 网络训练细节

数据处理. 首先将深度图转化到相机坐标系下，使用手的中心点截取边长为300 mm的立方体，再转化为128×128的深度图，对深度值进行归一化后输入网络. 在虚拟数据集中，手的中心点由21个关节点的平均坐标计算得到；在真实数据集中，为了与其他方法进行公平对比，使用Moon等[11]的方法中单独训练得到的手心坐标截取手部. 训练时对数据集进行随机旋转、平移、缩放.

训练流程. 网络首先在虚拟数据集上进行预训练，预训练时对特征提取网络输出的特征图和关节点、手模型输出的778个顶点和关节点进行监督 $\text{Loss}=L_{\text{dense}}+L_{\text{jt}}+L_{\text{verts}}+L_{\text{coll}}+L_{\text{region}}$. 在真实数据集上优化时，使用局部深度损失和碰撞损失进行自监督训练 $\text{Loss}=L_{\text{coll}}+L_{\text{region}}$. 注意在训练完成后，推理阶段仅保留特征提取网络和3D关节点估计分支（图2中蓝色箭头连接部分）

使用Adam优化器进行训练，使用初始学习率 10^{-4}，10^{-6} 分别训练40轮和25轮，当网络表现到达瓶颈时降低学习率为原来的0.1倍.

4 实验

4.1 数据集与评价标准

虚拟数据集 本文使用MANO手模型[24]与3D渲染技术，生成16 000的训练数据与1 600的测试数据，标注数据包括21个关节点坐标与手部网格778个顶点的坐标. 使用该数据集对网络进行预训练，使网络具备一定的特征提取能力.

NYU手部姿态数据集[13] 该数据集是用三个同

步深度摄像头从3个方向采集的,本文中仅使用从正面采集的数据(View 1),包含72 757帧训练数据以及8 252帧测试数据,每一帧包含36个关节点标注信息,本文选取其中21个与MANO手模型[24]定义最为相近的21个关节点进行误差评估.

本文使用两个评价标准对模型的精确度与鲁棒性进行评估.(1)测试集中单个/所有关节点的平均关节点误差.关节点误差指预测的关节点与标注之间的欧氏距离(单位为mm).(2)合格帧的数量占整体测试集的百分比.当单帧中每个关节点的误差都小于一定阈值时,视该帧合格.

4.2　实验结果

超参选择　分别对局部深度一致性损失中所使用超参数的不同取值进行实验,选择最优参数.如表2所示,前两行是局部深度的范围对平均关节点误差的影响.局部深度的范围由式(6)中高斯分布的标准差σ决定,σ越大,局部深度区域越大.可见平均关节点误差随着σ增大先下降、后上升,且σ由1.5增加到2.0,相比σ由1.5降到1.0,关节点误差增加幅度更大,说明局部区域大到一定程度时,网络效果下降明显.当$\sigma=1.5$时效果最好.表2后两行表示局部区域个数N(式(8))对网络效果的影响.分别设置局部区域个数为6(掌心及各手指指根)、11(掌心及各手指指根、第一指节)、16(掌心及各手指除指尖外的节点)、21(所有关节点).整体而言,N的影响相比σ较小.当$N=16$时,局部区域基本覆盖手的主要关节点,与使用全部关节点效果相差无几.本文之后的实验取使网络效果最好的参数$\sigma=1.5$, $N=21$.

表2　不同超参取值下网络的平均关节点误差

σ	1.0	1.5	2.0	2.5
平均关节点误差	16.17 mm	15.88 mm	16.84 mm	17.71 mm
N	6	11	16	21
平均关节点误差	16.46 mm	16.17 mm	15.90 mm	15.88 mm

训练数据对网络效果的影响　表3所示是本文所提方法在不同数据集上测试的效果.第1行表示模型在虚拟数据集的训练集上预训练(pretrained)后,在虚拟测试集上的平均关节点误差,为4.67 mm,说明模型对虚拟数据集的拟合效果良好.表中第2~5行实验均为在真实数据集上测试的结果.第2行表示不拟合真实数据、直接测试预训练模型在真实测试集上的表现,误差上升到28.74 mm,说明虚拟数据与真实数据之间的领域误差较大,使用现有表现较好的有监督网络AWR[8],在虚拟数据集上预训练的模型也无法直接应用到实际场景中.由于模型在真实数据集上拟合时不需要监督信息,因此可以将训练集、测试集进行组合以优化网络.第3~5行表示在不同的真实数据组合上拟合预训练模型的效果.与有监督方法不同,直接在测试集上训练网络,并没有因为网络"见过"数据而有更好的表现(第3行),而是数据量越多,网络对真实数据的适应能力越强,将训练集与测试集结合起来训练网络效果最好(第5行).

NYU数据集与MANO手模型关节点定义偏差　由于本文网络是在由MANO手模型渲染的虚拟数据集上预训练的,预测的关节点坐标服从手模型的定义,与NYU数据集的标注之间存在偏差.Endri等[7]直接将整体关节点误差减去最小关节点误差以消除这种偏差,然而这种方式无法证明整体关节点的偏移方向与最小关节点误差是一致的.为了准确计算偏差的大小以及考虑这种偏差对本文所提方法的效果的影响,使用MANO手模型和LM(Levenberg-Marquardt)迭代优化算法对NYU数据集的关节点坐标进行拟合,得到基于MANO手模型定义下对应的NYU数据集的关节点坐标.二者对比如图3所示,红线表示MANO,蓝线表示NYU原始数据集的关节点标注.可以看出偏差主要体现在两个方面.(1)关节点定义不同.MANO手模型上手指节点之间是等距的,而NYU数据集指尖节点与其父节点的距离略小于其他指间节点的距离.另外从图3中可以明显看出NYU定义的手腕节点相比MANO更靠近大拇指.(2)NYU关节点标注带有随机误差.MANO的关节点是在手的网格蒙皮中的固定位置,而NYU的关节点由于人工标注带有随机误差,在手上的位置并不是严格固定的.如图3绿圈的位置,NYU定义的关节点在MANO定义的关节点前后摆动.

表3　不同数据集上训练的网络的平均关节点误差

序号	方法	平均关节点误差/mm
1	pretrained (test on Sync.)	4.67
2	pretrained	28.74
3	finetune on test	16.55
4	finetune on train (ours)	15.88
5	finetune on train+test	15.31

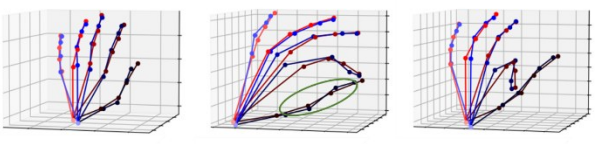
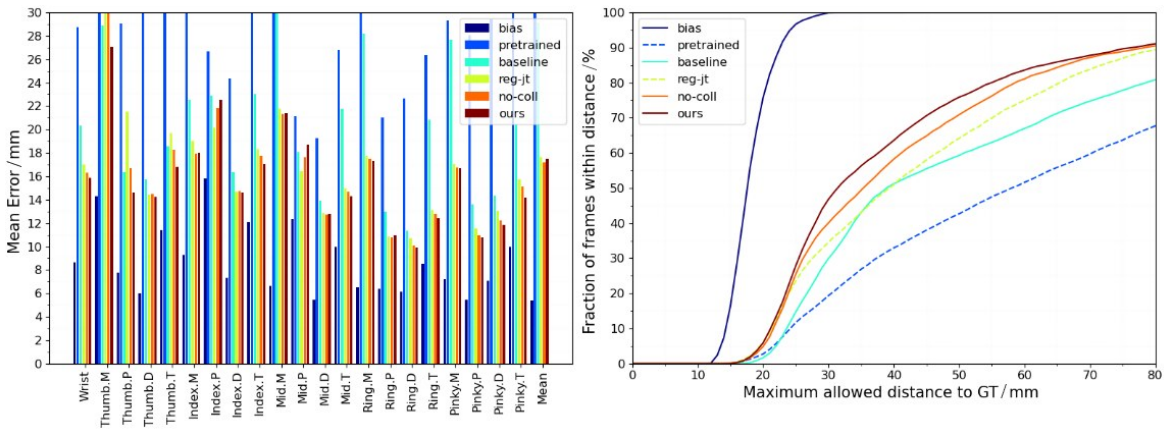


图3　NYU数据集(蓝线)与MANO模型(红线)关节点定义对比

消融实验 为了证明本文提出的局部深度一致性损失的有效性，设计了不同网络结构的消融实验，表4所示是不同方法的平均关节点误差. 其中bias指使用LM算法拟合NYU数据集得到的基于MANO手模型的关节点坐标估计结果，用于计算两种关节点定义之间的偏差；pretrained表示在虚拟数据集上训练、但未在真实数据上进行拟合的预训练模型；baseline指直接对整个输入与输出深度图计算损失；no-coll用于评估碰撞损失的作用；reg-jt和ours分别指利用手模型参数回归(REG)分支和3D关节点坐标估计(AWR)分支得到的关节点估计结果对深度图进行局部区域截取.

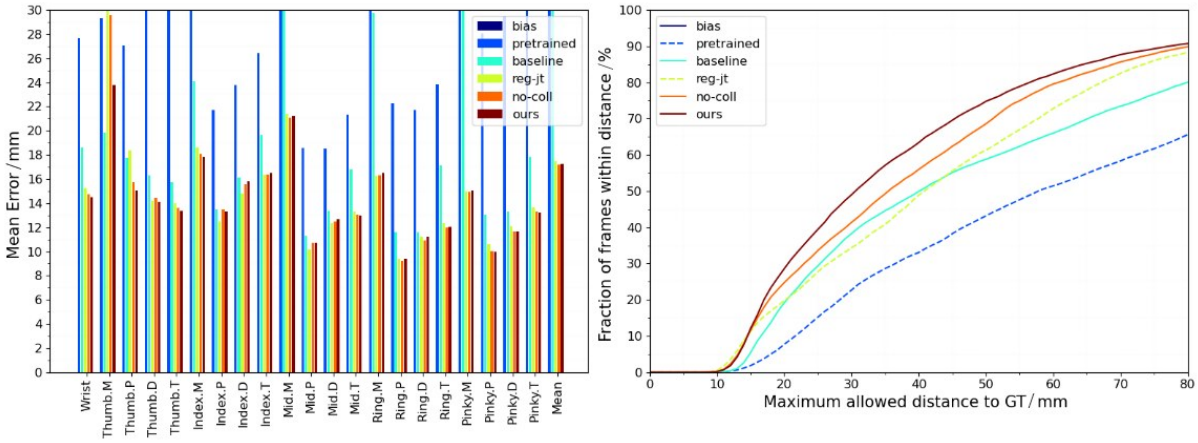
由表4可以看出，直接使用预训练模型在真实数据上进行测试存在较大误差(第2行，28.74 mm)，说明虚拟数据与真实数据之间存在较大的领域误差；baseline直接拟合完整的输入输出数据进行优化，效果提升了8.4 mm，验证了在真实数据上进行优化的必要性；reg-jt使用REG网络估计的MANO关节点对输入输出深度图进行截取并计算局部深度损失，对关节点附近的深度进行针对性的优化后，性能进一步提升了3.37 mm(第5行)；使用AWR网络生成的关节点坐标进行局部深度的截取，网络平均关节点误差再次下降1.09 mm(第6行)，由于局部深度的截取对初始姿态估计要求较高，而手模型参数的估计是高度非线性任务，对变化比较敏感，因此AWR输出的关节点坐标相比REG较为稳定，用于局部深度损失的计算中效果较好. 从表中还可以看出使用碰撞损失为网络带来了0.44 mm的提升(第4行)，由于手模型本身带有较强的约束与先验知识，碰撞损失能带来的效果提升比较有限. 使用MANO标注信息进行评估，两种关节点定义本身的固有偏差约为8.63 mm(第1行). 纵向对比不同方法之间的变化趋势大致相同，横向对比去除关节点定义的偏差后，整体效果提升了大约2mm.

图4和图5是对应方法在NYU，MANO两种关节点定义下的关节点误差与合格帧占比示意图. 从图4中



(*a*) 平均关节点误差 (*b*) 合格帧占比

图4 NYU标注下不同方法的平均关节点与单个关节点误差和合格帧占比



(*a*) 平均关节点误差 (*b*) 合格帧占比

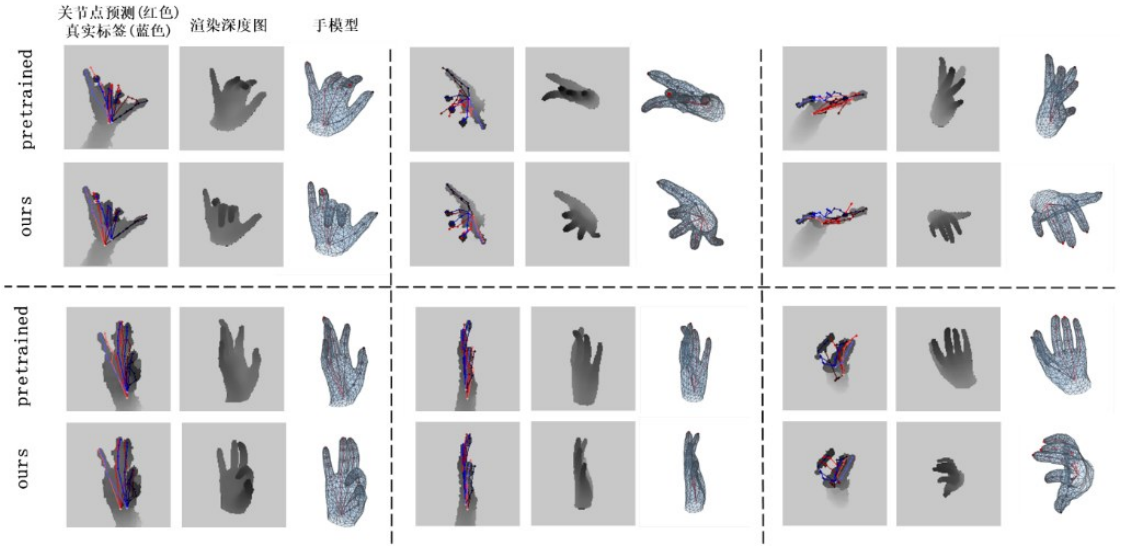
图5 MANO标注下不同方法的平均关节点与单个关节点误差和合格帧占比

表4　不同方法的平均关节点误差

序号	方法	平均关节点误差(NYU)/mm	平均关节点误差(MANO)/mm
1	bias	8.63	0
2	pretrained	28.74	27.67
3	baseline	20.34	18.61
4	no-coll	16.32	14.75
5	reg-jt	16.97	15.22
6	ours	15.88	14.46

左侧可以看出,单个关节点误差最大的是Thumb.M(约27 mm),去掉关节点定义的偏差后,由图5左侧可以看出,Thumb.M的关节点误差下降了3 mm,不同关节点之间的误差差距也略有减小. 当阈值为20 mm时,合格帧占比由不足10%提升到30%. 阈值越大,关节点定义偏差的影响越小,说明考虑关节点定义偏差可以纠正估计大致准确的精度,但对于估计失败的样本失去校正功能. 此外,由图5还可看出,模型估计掌心节点([Finger]. M)的误差要大于指间节点,这是由于不同掌心节点的局部深度相差较小,网络较难提取不同节点之间的特征信息,并且在大部分遮挡情况下,掌心节点尤其是大拇指的根节点(Index. M)被遮挡的概率较大(例如握拳动作),进一步增加了网络估计的难度.

可视化分析　由于自监督任务是模型拟合任务,优化网络的能量函数与上述两个评价指标只在一定程度上呈正相关,但不能全面评估网络的效果,因此对关节点估计结果以及深度图拟合效果进行可视化分析. 如图6所示共6组结果,每组中第1列是输入深度图、网络预测的关节点(红色)与真实关节点标注(蓝色),第2列是网络渲染的深度图,第3列是手的网格结构及对应关节点;每组中第1行是预训练模型直接在真实数据上测试的结果,第2行是本文所提方法的效果. 从图中可以看出,本文所提方法在网络的初始预测稍有偏差的情况下,能够纠正网络估计误差,且对手指、手掌遮挡、极端视角造成的部分深度值缺失等情况具有良好的适应能力. 然而,当深度图的深度值存在大量缺失时,网络失去初始预测偏能力时,纠正能力也相应减弱.



(a) 无遮挡　(b) 少部分遮挡　(c) 大部分遮挡

图6　预训练模型(pretrained)与本文方法(ours)的可视化对比

红线: NYU数据集标注. 蓝线: 网络预测结果

与其他方法对比　图7与表5所示是本文方法(基于NYU和MANO标注信息的评估结果)与其他方法的合格帧占比、平均关节点误差对比. 其中Refine-3D[15], Wan CD[16]与DualGridNet[14]与本文方法都是自监督方法,后两种方法使用了多视角的信息,公平起见,本文只与其方法中的单视角结果对比;DeepPrior[26]、DeepModel[27]和Feedback[23]是有监督方法. 从表5中可以看出,本文所提方法无论是与有监督方法还是无监督方法相比,平均关节点误差表现都较好;从图7中可以看出,本文所提方法在阈值小于25 mm时,合格帧占比优于其他方法,而当阈值增加时,表现略逊色于其他方法,说明本文方法对NYU数据集中的部分极端视角数据失去预测能力,导致关节点误差较大. 当遇到极端视角、深度值缺失情况较严重时,网络失去初始预测能力,从而无法纠正错误的初始关节点估计结果. 因而需要确保网络基本的预测能力,尽可能保证虚拟数据的分布的多样性与全面性,保证虚拟数据集和真实数据集的相似程度,如向虚拟数据中加入随机噪声、深度值缺失,或者利用GAN处理虚拟数据集,使其风格趋近真实数据等.

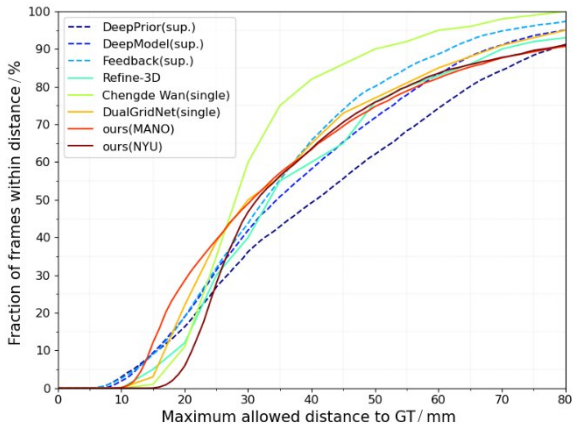


图7 有监督方法(虚线)和自监督方法(实线)的合格帧占比对比

表5 与有监督方法(sup.)和自监督方法的平均关节点误差对比

方法	平均关节点误差/mm
DeepPrior(sup.)	20.75
DeepModel(sup.)	17.04
Feedback(sup.)	15.97
Wan CD(single)	17.79
DualGridNet(single)	16.96
ours(MANO)	15.88
ours(NYU)	14.46

5 结束语

本文提出局部深度一致性损失以解决自监督3D手部姿态估计中直接拟合完整深度图带来的特征提取困难、模型难以对齐的问题. 通过使用3D关节点坐标估计分支(AWR)的估计结果,将输入、输出深度图显式地解耦合为围绕关节点的不同区域,通过学习各区域内深度值的一致性,使网络有针对性地学习各个关节点对应区域的特征,削弱虚拟数据与真实数据的领域误差带来的全局影响. 在NYU数据集上,充分的消融实验、可视化结果以及与其他方法的对比,证明了所提方法的有效性与可扩展性.

参考文献

[1] 任海兵,祝远新,徐光祐,等.基于视觉手势识别的研究-综述[J]. 电子学报, 2000, 28(2): 118-121.
REN H B, ZHU Y X, XU G Y, et al. Vision-based recognition of hand gestures: A survey[J]. Acta Electronica Sinica, 2000, 28(2): 118-121. (in Chinese)

[2] 管业鹏.复杂人机交互场景下的指势用户对象识别[J].电子学报, 2014, 42(11): 2135-2141.
GUAN Y P. Pointing user recognition in human-computer interaction with cluttered scene [J]. Acta Electronica Sinica, 2014, 42(11): 2135-2141. (in Chinese)

[3] 徐一华,李善青,贾云得.一种基于视觉的手指屏幕交互方法[J]. 电子学报, 2007, 35(11): 2236-2240.
XU Yi-hua, LI Shan-qing, JIA Yun-de. A vision-based method for finger-screen interaction[J]. Acta Electronica Sinica, 2007, 35(11): 2236-2240. (in Chinese)

[4] 武汇岳,王建民,戴国忠.基于小样本学习的3D动态视觉手势个性化交互方法[J]. 电子学报, 2013, 41(11): 2230-2236.
WU HUI-YUE, WANG JIAN-MIN, DAI GUO-ZHONG. Personalized interaction techniques of vision-based 3D dynamic gestures based on small sample learning[J]. Acta Electronica Sinica, 2013, 41(11): 2230-2236. (in Chinese)

[5] CUI J, KUIJPER A, SOURIN A. Exploration of natural free-hand interaction for shape modeling using leap motion controller [C]//Proceedings of the International Conference on Cyberworlds(CW). Chongqing: IEEE Computer Society, 2016: 41-48.

[6] 齐静, 徐坤, 丁希仑. 机器人视觉手势交互技术研究进展[J]. 机器人, 2017, 39(4): 565-584.
QI J, XU K, DING X L. Vision-based hand gesture recognition for human-robot interaction: A review [J]. Robot, 2017, 39(4): 565-584. (in Chinese)

[7] WAN C D, PROBST T, GOOL L V, et al. Dense 3d regression for hand pose estimation[C]//Computer Vision and Pattern Recognition (CVPR). Utah: Computer Vision Foundation / IEEE Computer Society, 2018: 5147-5156.

[8] HUANG W T, REN P F, WANG J Y, et al. Awr: adaptive weighting regression for 3d hand pose estimation[C]//Association for the Advancement of Artificial Intelligence (AAAI). New York: Journal of Artificial Intelligence Research, 2020: 11061-11068.

[9] CHEN Y J, TU Z G, GE L H, et al. SO-handnet: self-organizing network for 3d hand pose estimation with semi-supervised learning[C]//International Conference on Computer Vision (ICCV). Seoul: IEEE. 2019: 6960-6969.

[10] GE L H, LIANG H, YUAN J S, et al. Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns[C]//Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE Computer Society, 2016: 3593-3601.

[11] MOON G, CHANG J Y, LEE K M. V2v-posenet: voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map[C]//Computer Vision and Pattern Recognition (CVPR) . Utah: IEEE Computer Society, 2018: 5079-5088.

[12] YUAN S X, YE Q, STENGER B, et al. Bighand2.2m benchmark: hand pose dataset and state of the art analysis [C]//Computer Vision and Pattern Recognition (CVPR) . Hawaii: IEEE Computer Society, 2017: 2605-2613.

[13] TOMPSON J, STEIN M, YANN L C, et al. Real-time continuous pose recovery of human hands using convolutional networks[J]. ACM Transactions on Graphics

(TOG), 2014, 169(33): 1-10.
[14] WAN C D, PROBST T, GOOL L V, et al. Dual grid net: Hand mesh vertex regression from single depth maps[C]// European Conference on Computer Vision (ECCV). Glasgow: Springer, 2020: 442-459.
[15] DIBRA E, WOLF T, ÖZTIRELI C, et al. How to refine 3d hand pose estimation from unlabelled depth data[C]//International Conference on 3D Vision. Qing Dao: Institute of Electrical and Electronics Engineers, 2017: 135-144.
[16] WAN CD, PROBST T, GOOL LV, et al. Self-supervised 3d hand pose estimation through training by fitting[C]// Computer Vision and Pattern Recognition (CVPR) . Long Beach: IEEE Computer Society, 2019: 10853-10862.
[17] MELAX S, KESELMAN L, ORSTEN S. Dynamics based 3d skeletal hand tracking[C]// Proceedings of Graphics Interface 2013. Toronto: Canadian Information Processing Society, 2013: 63-70.
[18] SINHA A, CHOI C, RAMANI K. Deephand: Robust hand pose estimation by completing a matrix imputed with deep features[C]//Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE Computer Society, 2016: 4150-4158.
[19] ZHANG H, BO Z H, YONG J H, et al. InteractionFusion: real-time reconstruction of hand poses and deformable objects in hand-object interactions[J]. ACM Transactions on Graphics, 2019, 38(4): 1-11.
[20] SUPANCIC III JS, ROGEZ G, YANG Y, et al. Depth-based hand pose estimation: Methods, data, and challenges[J]. International Journal of Computer Vision, 2018, 126 (11): 1180-1198.
[21] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE Computer Society, 2016: 770-778.
[22] ZHANG X, LI Q, MO H, et al. End-to-end hand mesh recovery from a monocular RGB image[C]//International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019: 2354-2364.
[23] OBERWEGER M, WOHLHART P, LEPETIT V. Training a feedback loop for hand pose estimation[C]//International Conference on Computer Vision (ICCV) . Santiago: IEEE Computer Society, 2015: 3316-3324.
[24] ROMERO J, TZIONAS D, BLACK MJ, hands Embodied: Modeling and capturing hands and bodies together [J]. ACM Transactions on Graphics (TOG). 2017, 36(6): 245:1-245:17.
[25] REN P F, SUN H F, HUANG W T, et al. Spatial-aware stacked regression network for real-time 3D hand pose estimation [J]. Neurocomputing, 2021, 437: 42-57.
[26] OBERWEGER M, WOHLHART P, LEPETIT V. Hands deep in deep learning for hand pose estimation[C]//Computer Vision Winter Workshop. Styria: Slovenian Pattern Recognition Society, 2015: 1-10.
[27] ZHOU X Y, WAN Q F, ZHANG W, et al. Model-based deep hand pose estimation[C]//International Joint Conference on Artificial Intelligence (IJCAI). New York: Margan Kaufmann, 2016: 2421-2427.

作者简介

王敬宇　男,1978年2月出生于吉林省长春市. 北京邮电大学网络与交换国家重点实验室教授、博士生导师. 研究方向为智能网络、人工智能、云计算、多媒体通信、多路径传输、流量工程等.
E-mail: wangjingyu@bupt.edu.cn

黄伟亭　女,1997年3月出生于福建省三明市. 北京邮电大学网络与交换国家重点实验室硕士生. 研究方向为人工智能、计算机视觉等.
E-mail: hwt97@bupt.edu.cn

刘　聪　男,1980年出生于山东省泰安市. 中国移动研究院高级工程师. 研究方向为人工智能、物联网等.
E-mail: liucong@chinamobile.com

戚　琦(通讯作者)　女,1982年6月出生于河北省廊坊市. 北京邮电大学网络与交换国家重点实验室副教授、博士生导师. 研究方向为智能边缘计算、轻量级神经网络、业务网络智能化等.
E-mail: qiqi8266@bupt.edu.cn

孙海峰　男,1989年出生于天津市. 北京邮电大学网络与交换国家重点实验室讲师、硕士生导师. 研究方向为人工智能、机器视觉、自然语言处理、深度学习等.
E-mail: hfsun@bupt.edu.cn

廖建新　男,1965年出生于四川省宜宾市. 北京邮电大学长江学者特聘教授、博士生导师. 研究方向为移动通信网络、业务网络化、人工智能、多媒体业务等. 中国电子学会会员编号:E190027234S.
E-mail: liaojianxin@ebupt.com