第 42 卷 第 19 期/2022 年 10 月/光学学报 研究论文

光学学报

基于手指点加强和多级特征融合的三维人手姿态估计

张开宜[1,2]，洪儒[1,2]，盖绍彦[1,2]，达飞鹏[1,2,3*]

[1]东南大学自动化学院，江苏 南京 210096；

[2]东南大学复杂工程系统测量与控制教育部重点实验室，江苏 南京 210096；

[3]东南大学深圳研究院，广东 深圳 518036

摘要 针对现有的三维(3D)人手姿态估计算法没有充分挖掘手指特性和关键特征作用的问题，提出了手指点加强(FPR)策略和多级融合注意力(MFSE)模块。FPR策略突出了人手点云中手指位置点的作用，加强了网络特征提取层对点云中手指位置点的关注，提高了手指关节点的回归精度。MFSE模块提高了分层网络提取和表达局部特征的能力，该模块实现了分层网络之间不同层次特征的融合和权重分配，增强了模型的鲁棒性和人手姿态估计的准确度。在两个公共基准数据集MSRA和ICVL上的实验表明，所提算法能够实现高精度的3D人手姿态估计。

关键词 机器视觉；三维点云；深度学习；注意力机制；手部姿态估计

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/AOS202242.1915001

Three-Dimensional Human Hand Pose Estimation Based on Finger-Point Reinforcement and Multi-Level Feature Fusion

Zhang Kaiyi[1,2], Hong Ru[1,2], Gai Shaoyan[1,2], Da Feipeng[1,2,3*]

[1]*School of Automation, Southeast University, Nanjing* 210096, *Jiangsu, China*;

[2]*Key Laboratory of Measurement and Control of Complex Engineering Systems, Ministry of Education, Southeast University, Nanjing* 210096, *Jiangsu, China*;

[3]*Shenzhen Research Institute, Southeast University, Shenzhen* 518036, *Guangdong, China*

Abstract The existing three-dimensional (3D) human hand pose estimation algorithms do not fully exploit the characteristics of fingers and the key features. To solve this problem, a finger-point reinforcement (FPR) strategy and a multi-layer fusion squeeze and excitation (MFSE) block are proposed. The FPR strategy highlights the role of the finger position points in the human hand point cloud, strengthens the attention of network feature extraction layers to the finger position points in the point cloud, and improves the regression accuracy of the finger joint points. The MFSE block improves the ability of the layered network to extract and express local features. This module realizes the fusion and weight distribution of different levels of features between the layered networks, thereby enhancing the robustness of the model and the accuracy of human hand pose estimation. Experiments on two public benchmark datasets, MSRA and ICVL, verify that the proposed algorithm can achieve high-precision 3D human hand pose estimation.

Key words machine vision; three-dimensional point cloud; deep learning; attention mechanism; hand pose estimation

1 引 言

近年来，手部姿态估计技术在很多领域(机器人、游戏、医学和汽车等)中都有大量的应用，如何快速且准确地获取手部姿态逐渐成为一个重要的研究方向[1-5]。由于三维(3D)人手姿态存在手指自相似度高和自遮挡严重等问题，故如何鲁棒地获取3D人手姿态的估计结果仍然存在一些挑战。在3D人手姿态估计

收稿日期：2022-01-19；**修回日期**：2022-03-21；**录用日期**：2022-04-16

基金项目：国家自然科学基金(51475092)、江苏省前沿引领技术基础研究专项(BK20192004C)、深圳市科技创新委员会(JCYJ20180306174455080)

通信作者：*dafp@seu.edu.cn

领域中，基于二维图像的方法取得了不错的成果，能估计出一定精度的3D空间内的人手姿态，满足部分实际应用的需要。然而，通过二维图像对3D空间中的手势进行推测必然存在失真的情况，故难以实现高精度的3D人手估计。3D数据极大程度上弥补了许多应用场景中二维图像所缺失的空间结构信息。目前的研究依据不同3D数据表示方式又可以分为基于深度图的方法[6-9]、基于体素的方法[10-11]和基于点云的方法[12-14]。

在各种表达方式中，点云表达方式简单，且更接近物体的原始立体表征，将3D点云作为输入避免了利用多视图间接推测获得关键点3D坐标的精度损失，也避免了体素这种规则化数据产生不必要的体积划分和影响数据的不变性。虽然点云数据受采集设备和坐标系影响，呈现出了一定的无序性，但是文献[15]和文献[16]提出的PointNet++网络很好地解决了这个问题，文献[12]也在此基础上提出了利用分层点云网络直接使用3D点云作为输入估计3D人手姿态的方法。虽然PointNet++网络可以稳定地提取点云特征，但点云数据潜藏了丰富的特征信息，故如何在网络模型的特征提取过程中加强对手指部位特征的提取是一个重要的问题。

针对上述问题，本文提出了手指点加强（FPR）策略和多级融合注意力（MFSE）模块。FPR策略提出了置信度参数以估计点云中某个点位于手指位置的可能性，再强化高置信度参数点在整个模型训练中的作用。人手姿态估计主要是回归人手指关节点的位置，而FPR策略能有效判断出对应点在手指位置处的置信度，加强网络特征提取阶段对手指点的关注，提高这些点提取的局部特征在整个网络中的权重，同时能抑制手掌位置处的点、手腕位置处的点和一些噪声点对网络模型学习过程的干扰，因此该策略能有效地提高人手关节点回归的准确性。MFSE模块结合了注意力机制[17]和残差网络[18]，其能够将有效的局部特征继承到更高维度的特征中，提高有效局部特征在网络中的作用，以此来强化分层点云网络对局部特征的提取与表达能力，聚焦更利于回归手部关节点的核心骨干特征。该方法能够有效地对分层网络中各通道之间的关系进行自适应权重分配并提高优秀的局部特征对全局特征的影响。在标准公共数据集MSRA[19]和ICVL[20]中的测试结果都证实了所提算法的有效性。

2 基本原理

所提算法以3D点云作为输入，输出M个关节的3D坐标。主要流程如图1所示，其中$N_1\times D_1$为N_1个具有D_1维特征的点，$N_2\times D_2$为N_2个具有D_2维特征的点。首先将点云通过最远点采样（FPS）法[12]降采样为N（$N=1024$）个点$[p_i\in\mathbf{R}^3, i=1,2,\cdots,N]$，继而在定向边界盒（OBB）中对采样点进行归一化[12]，并计算出归一化点坐标(x,y,z)、法向量$\boldsymbol{n}=(n_x,n_y,n_z)$和FPR策略中的置信度参数$P$，构成$D(D=7)$个特征。然后，将上述的$N\times D$维特征作为分层点云网络的输入，并在每两个特征提取层之间加入所提的MFSE模块，网络输入经过三个特征提取层后可得到全局特征。最后，将全局特征输入到全连接层中，得到M个关节的3D坐标。该算法以端到端的形式完成了高精度手部姿态的估计任务。

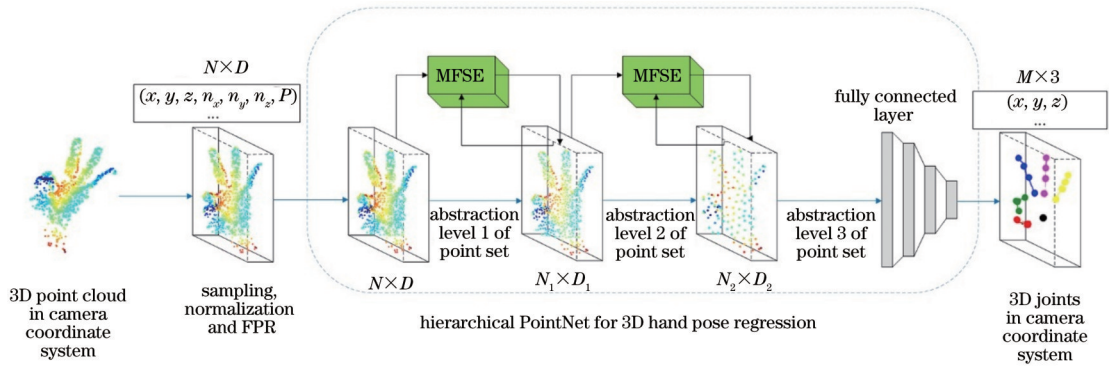

图1 总体算法流程概览

Fig. 1 Overview of overall algorithm flow

2.1 手指点强化策略

在3D手部姿态估计任务中，手指关节点的估计是最重要的内容。在标准公共数据集MSRA和ICVL的关节点真值中，手指点关节的数量占据所有关节点数量的90%以上。现实中手掌所占有的面积远远超过手指，这也体现在了点云的分布上。手掌部分占据了点云中的大部分点，但这些点对于手指关节点的回归没有太大的用处。若能够在3D点云中直接确定手指部位的点，提高网络对这些点的特征提取能力，则能够提高学习模型回归的3D手势估计精度。然而，由于人手的高自由度和点云的无序性，故难以在人手点云中直接寻找到手指位置的点。

针对这个问题，本文提出了FPR策略，该策略提出了置信度参数来描述点云中对应点在手指位置的可能性，该参数体现了同一个点附近不同范围内点组成的两个平面之间的夹角大小，置信度参数越大，夹角越大。手指作为一个圆柱体，且能够变化的姿势非常多，相较于手掌这个近乎于平面的部位而言，点云中该位置附近点的置信度参数的值都比较大。因此，点云中某点置信度参数的值越大，其在手指位置的可能性也

越高。如图 2 所示，对 FPS 方法[16]获得的 1024 个采样点使用 FPR 策略，将置信度计算结果最高的 128 个点标黑突出，可以看出这 128 个标黑的点基本上都在手指位置附近，这说明置信度参数的提出是合理的。

图 2　3组不同手势的对比图(上方标黑的点为置信度参数较高的部分，下方为原始点云)

Fig. 2　Comparison of three groups of gestures (black dots on top are parts with higher confidence parameters, and original point clouds are shown on bottom)

下面将介绍本文置信度参数的具体实现步骤，所有的计算过程都是在归一化后的点云中完成的。首先是计算对应点的法向量，对于点云中的任一点，在一定范围内找到该点附近 K_1 个点，如不足 K_1 个点就只取范围内的点，并将其置信度参数 P 置 0，拟合这些选中点组成一个平面，再计算这个平面的法向量 $\boldsymbol{n}_1$。同上，再计算该点附近 K_2（$K_2>K_1$）个点对应的法向量 $\boldsymbol{n}_2$。法向量 $\boldsymbol{n}_1$ 和 $\boldsymbol{n}_2$ 的表达式为

$$\boldsymbol{n}_1=\min_{\|\boldsymbol{n}\|_2=1}\sum_{i=1}^{K_1}\left[(\boldsymbol{x}_i-\boldsymbol{c})^{\mathrm{T}}\boldsymbol{n}\right]^2, \tag{1}$$

$$\boldsymbol{n}_2=\min_{\|\boldsymbol{n}\|_2=1}\sum_{i=1}^{K_2}\left[(\boldsymbol{x}_i-\boldsymbol{c})^{\mathrm{T}}\boldsymbol{n}\right]^2, \tag{2}$$

式(1)计算出对应点 $\boldsymbol{c}\in\mathbf{R}^3$ 附近的点 $\boldsymbol{x}_i\in\mathbf{R}^3(i=1,2,\cdots,K_1)$ 拟合的平面对应的单位法向量 $\boldsymbol{n}_1=(n_{x1},n_{y1},n_{z1})\in\mathbf{R}^3$。式(2)计算出对应点 $\boldsymbol{c}$ 附近的点 $\boldsymbol{x}_i\in\mathbf{R}^3(i=1,2,\cdots,K_2)$ 拟合的平面对应的单位法向量 $\boldsymbol{n}_2=(n_{x2},n_{y2},n_{z2})\in\mathbf{R}^3$。

其次，计算出上述两个法向量之间的欧几里得距离，放大这个欧几里得距离得到的就是置信度参数。最后，为了进一步强化手指位置附近的点对后续特征提取的作用，在置信度参数值最高的 128 个点附近范围内找出置信度参数值低于阈值的点，FPR 策略认为这些点被错误地低估了，将这些点的置信度参数值修改为阈值。置信度参数的计算方法为

$$\theta=(n_{x1}-n_{x2})^2+(n_{y1}-n_{y2})^2+(n_{z1}-n_{z2})^2, \tag{3}$$

$$P=k\theta, \tag{4}$$

式中：θ 为上述两个法向量之间的欧氏距离；k 为放大系数。由于 θ 的值过小，故需要乘以参数 k 将其放大，这样就得到了置信度参数 P，具体参数的选取在实验细节部分进行介绍。

在实际计算中能够确定在加入了 FPR 策略后，点云中手指位置的点相较其余位置有较高的置信度参数值。强化网络模型对这些拥有高置信度参数值的点的特征提取，提高这些点提取出的局部特征的权重，就能够提高网络模型估计的 3D 人手姿态的精度。此外，对于手腕位置的点和一些噪声点，由于手腕位置点分布的曲面非常平缓，故手腕点对应的法向量与拟合平面的法向量之间的欧氏距离较小，此时计算得到的置信度参数偏低，而噪声点一般游离于主体点云中，其置信度易被 FPR 策略置 0。因此，FPR 策略降低了噪声点对网络模型学习过程的干扰，提高了模型的鲁棒性。

2.2　多级融合注意力模块

神经网络模型的特征提取能力可以通过增加一些嵌入式的模块来提高，从而降低对标注信息的要求，提高模型的泛化能力。挤压激励(SE)网络是一个能够自适应地调整各个通道间权重的模块，该模块在二维图像任务中表现良好，对于许多任务的效果都有显著提升[17]。然而，无序的点云数据与二维图像相比，数据维度较高，复杂性更强，故在点云网络中直接加入 SE 模块对模型的效果提升有限。

针对上述问题，本文提出了 MFSE 模块，MFSE 模块结合了 SE 模块与残差连接，匹配了分层点云网络。此外，与 SE 模块相比，MFSE 模块不但自适应地调整了特征提取层中各个通道间的权重，而且加强了更有效的局部特征在分层次的特征提取过程中的影响力。

MFSE 的结构如图 3 所示，与文献[17]不同，对于点云网络，由于特征维度的区别，故其 Squeeze 部分也与二维图像不同。在图 3 中，B 为批训练数据大小，N_j 为样本在第 j 层的点个数，C_j 为样本在第 j 层的特征通道数，H 为二维图像的高，W 为二维图像的宽。将从上一特征提取层获得的点云中间特征与经过特征提取层和 Squeeze 操作的当前特征提取层的中间特征作为输入，再利用 Concat 操作拼接不同层次间的局部特征得到融合特征。最后，通过全连接层和归一化函数得到当前层级特征通道的自适应权值。该权值与每个特征通道对应的影响力呈正相关，MFSE 模块比较了不同层级中间特征的全局影响力，以此来融合两层不同特征通道间的显著特征，计算出对全局特征更有竞争力的局部特征，并提高这些更有效的局部特征对全局特征的贡献。

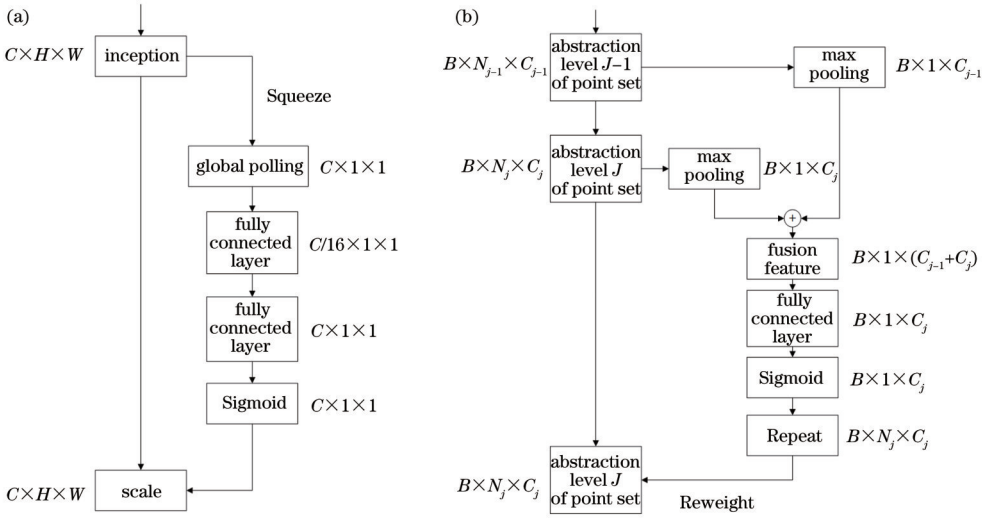

图 3 SE模块和MFSE模块。(a) SE模块;(b) MFSE模块

Fig. 3 SE module and MFSE module. (a) SE module; (b) MFSE module

MFSE子模块的两组输入分别为第$j-1$层的输出特征$f_{j-1}\in\mathbf{R}^{N_{j-1}\times C_{j-1}}$和经过当前特征提取层的第$j$层的输出特征$f_j\in\mathbf{R}^{N_j\times C_j}$。MFSE子模块可表示为

$$W_{\mathrm{MFSE}}(f_{j-1},f_j)=\varphi\left[F(f_{j-1},f_j)\right],\tag{5}$$

式中:$\varphi(\cdot)$为含有归一化函数的全连接层,本实验中采用Sigmoid激励函数;$F(\cdot)$的作用是抽象出最显著特征并进行拼接。$W_{\mathrm{MFSE}}\in\mathbf{R}^{1\times C_j}$为最终的权值,经过Repeat操作和Reweight操作后与f_j融合。不同层次特征的融合特征为

$$F(f_{j-1},f_j)=P(f_{j-1})\oplus P(f_j),\tag{6}$$

式中:$P(\cdot)$为全局特征聚合函数,如最大池化函数;$\oplus$为Concat操作,目的是拼接两组输出特征$P(f_{j-1})$和$P(f_j)$,得到的输出为$F(f_{j-1},f_j)\in\mathbf{R}^{1\times(C_{j-1}+C_j)}$。

3 实验、分析与讨论

3.1 实验数据集和评估标准

本文在公共手姿态数据集MSRA和ICVL上评估了所提算法。MSRA数据集有9个不同的主题,每个主题手的大小、胖瘦都不同,每个主题包含17个手势,每种手势的每一帧都有21个关节点真值,共计76000帧数据。ICVL数据集包含22000个训练帧和1600个测试帧。每一帧有16个关节点真值。

3D人手姿态估计的性能评估包括平均误差距离和良好帧比例[12]。平均误差距离能度量3D人手姿态回归的关节点值与真值之间的平均误差大小,该值越小,模型回归的精度越高。良好帧比例能描述误差小于一定阈值的测试帧占测试帧总数的比例,若该值在阈值越小时越大,则表明此时的模型越准确。

3.2 实验细节

本文所用计算机硬件配置:中央处理器(CPU)为Intel® Core™ i5-10600KF,主频为4.10 GHz;图形处理器(GPU)为GeForce RTX 3070。深度学习框架使用PyTorch 1.9.1,环境配置使用anaconda 4.10.1。

在MSRA数据集中,在8个主题上对所提模型进行训练,并在剩余的主题上进行测试,没有对此数据集进行任何数据增强。取5次训练结果的平均指标作为所提算法的结果。

在ICVL数据集中,使用随机森林策略(RDF)进行手部分割,并使用随机手臂长度和随机拉伸因子来增强训练数据。取5次训练结果的平均指标作为所提算法的结果。

本文用主成分分析(PCA)方法来对采样点云中对应点的最邻近点进行分析以得到拟合平面,并选取最邻近点的个数为30[12],即式(1)中的K_1为30。对于分层点云网络,保持与文献[16]相同的设置,使用FPS法对点云局部区域的质心进行采样,并用球状查询法对点进行分组,球状查询法的第一级半径设置为0.1,第二级半径设置为0.2。对于不同层特征提取层输出的特征维度,与文献[12]保持一致,即$N_1=512$,$N_2=128$。在训练过程中,本文使用的优化器是Adam,初始学习率为0.001,批次为24,正则化强度为0.0005。在训练了60轮后,将学习率变成初始学习率的1/10,总共在测试集上训练80轮。

对于所提的FPR策略,与Pointnet++相同,使用球状查询法寻找一定范围内的点,设定半径为0.15。当式(2)选取不同K_2时,其置信度参数对实验结果的影响也不同,本文选取$K_2=50$时的置信度参数作为实验的标准。式(4)的放大系数为$k=10$,最后的阈值选择的是0.45。

为了比较所提算法对3D人手姿态估计的作用,本文引用了一个基础的3D人手姿态估计网络作为Baseline[12]。该Baseline同样使用OBB方法[12]进行归一化,也通过式(1)计算出法向量(所有参数保持一致)。然后,利用PointNet++分层地提取全局特征,并利用全连接层回归估计关节点的3D坐标。

3.3 消融实验

3.3.1 手指点加强策略

为了验证FPR策略对整个3D手势估计任务的贡献，本文在MSRA数据集上比较了该策略对不同手指关节点回归精度的影响。

本文的实验设计了两个对照组，一是Baseline，二是Baseline＋FPR。在实验中，选取了11个关节点的估计值与真值的误差来代表FPR策略对不同关节点的影响，所选部位为食指、中指、无名指、小指和大拇指的指根和指尖部分（R代表指根，T代表指尖），以及手掌中心。各个关节点在不同方法下的误差距离如表1所示。

表1 FPR策略在MSRA数据集上各个关节的误差距离

Table 1 Error distance of each joint of FPR strategy on MSRA dataset unit: mm

Joint	Baseline	Baseline＋FPR
Palm	8.6665	8.2581
Index_R	6.8979	6.4721
Index_T	11.1541	10.5203
Mid_R	5.4392	5.1762
Mid_T	10.9231	10.3742
Ring_R	5.7619	5.5396
Ring_T	9.7697	9.4127
Pinky_R	7.4491	7.1871
Pinky_T	10.2487	9.6389
Thumb_R	7.5354	7.2211
Thumb_T	13.9441	13.7389
Mean error	8.5530	8.1930

比较这11个关节点在Baseline和Baseline＋FPR下不同关节的误差距离可知，FPR策略不仅对每个手指关节点的估计精度都有显著提升，还对掌心关节点坐标有明显提升，这表明FPR策略对不同位置的点进行可信度分析是有效的，不同的位置点对应不同范围的置信度，这些置信度可辅助网络在特征提取过程中对不同位置的点赋予不同的关注度，进而提升网络对关节点的估计精度。因此，FPR策略在提高3D人手姿态估计精度方面能发挥较大的作用。

从表1可知，FPR策略对于不同位置关节点的精度提升也不同：对于手掌关节点，精度提高了4.71%；对于食指的指根和指尖，精度分别提高了6.17%和5.68%；对于中指的指根和指尖，精度分别提高了4.84%和5.03%；对于无名指的指根和指尖，精度分别提高了3.86%和3.65%；对于小指的指根和指尖，精度分别提高了3.52%和5.95%；对于大拇指的指根和指尖，精度分别提高了4.17%和1.47%；所有关节点的平均误差提高了3.97%。

从表1还可以看出，除了大拇指这个形状较为特殊的手指之外，FPR策略对指尖关节点误差减少的幅度比指根关节点大，表明FPR策略对估计难度更高的指尖关节点也能起到很好的作用。因此，所提策略能减少关节点的估计误差，提高3D人手姿态估计任务的精度。

3.3.2 多级融合注意力模块

为了验证MFSE模块的有效性，本文选择了三个对照组，分别是Baseline、Baseline＋SE、Baseline＋MFSE，并比较了它们对不同关节估计精度的影响，如表2所示。

表2 不同方法在MSRA数据集上各个关节的误差距离

Table 2 Error distance of each joint on MSRA dataset by different methods unit: mm

Joint	Baseline	Baseline＋SE	Baseline＋MFSE
Palm	8.6665	8.4958	8.3172
Index_R	6.8979	6.5829	6.5038
Index_T	11.1541	10.6960	10.4891
Mid_R	5.4392	5.2140	5.1937
Mid_T	10.9231	10.2835	10.1443
Ring_R	5.7619	5.6199	5.5399
Ring_T	9.7697	9.4182	9.2983
Pinky_R	7.4491	7.3386	7.1515
Pinky_T	10.2487	10.1666	9.7651
Thumb_R	7.5354	7.4101	7.2387
Thumb_T	13.9441	13.9072	13.7871
Mean error	8.5320	8.3280	8.2050

在表2所示的三个对照组中，加入了MFSE模块的模型的表现远好于Baseline和Baseline ＋ SE的模型，故所提的MFSE模块对3D人手姿态估计任务能起到更好的作用。

从表2可知，与Baseline模型相比，MFSE模块对不同位置关节点的估计精度的提高也不同：对于手掌关节点，精度提高了4.03%；对于食指的指根和指尖，精度分别提高了5.71%和5.96%；对于中指的指根和指尖，精度分别提高了4.51%和7.12%；对于无名指的指根和指尖，精度分别提高了3.85%和4.83%；对于小指的指根和指尖，精度分别提高了4.00%和4.72%；对于大拇指的指根和指尖，精度分别提高了3.94%和1.13%；所有关节点的平均误差提高了3.83%。

从表2还可以看出，MFSE模块对于所有关节点的误差减小幅度都高于SE模块，尤其是在小指和大拇指指尖处，在SE模块在这两处的作用不明显的情况下，MFSE仍能有很好的表现，说明MFSE能够更好地提高有效特征对最终结果的贡献。

3.3.3 手指点加强和多级融合注意力共同作用分析

本文又设计了实验分析比较FPR和MFSE单独作用和共同作用时对3D人手姿态估计任务的精度提高。

由图4可知，当所提FPR方法和MFSE模块同时作用时，每个关节点的误差比单独使用更低，平均关节误差从8.532 mm降低到了7.942 mm，比单独加入

FPR 策略的情况(8.193 mm)和单独加入 MFSE 模块的情况(8.205 mm)都要好。

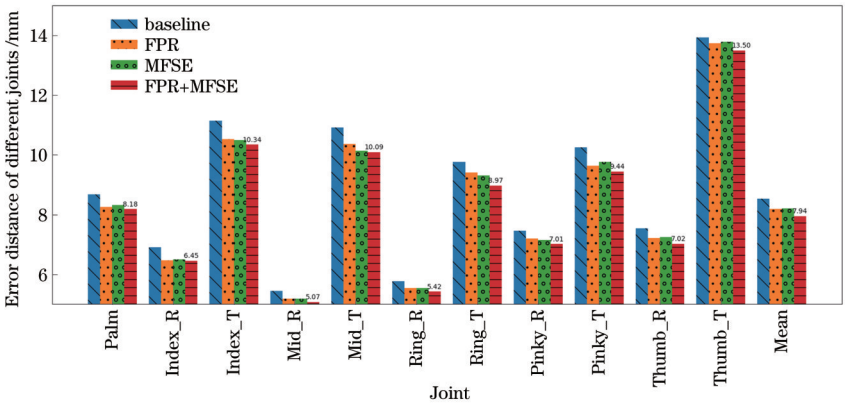

图 4　各方法手指关节点误差

Fig. 4　Error of each finger joint point by different methods

由图 5 可知:加入 FPR 方法和 MFSE 模块都提高了测试帧中低误差测试帧的比例;在 0～30 mm 范围内,两者单独作用时都提高了估计误差在这些误差范围内的测试帧数量;当两者共同作用时,网络模型在 0～50 mm 范围内的良好帧比例都有显著提高。

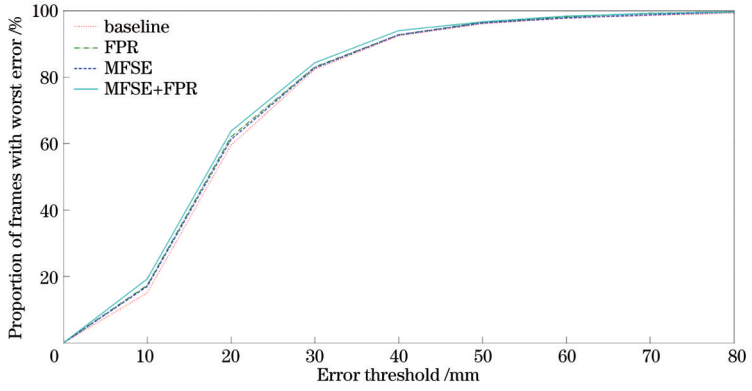

图 5　误差在不同阈值内的测试帧比例

Fig. 5　Proportion of test frames with errors within different thresholds

因此,FPR 方法和 MFSE 模块两者结合起来能够对 3D 手势估计任务的精度提高起到更好的作用,比两个方法单独作用的效果更好。

3.4　实验结果比较

所提算法与 8 个先进方法的比较如表 3 所示。在 MSRA 数据集上,所提算法较 Hand PointNet[12]、3D DenseNet[21]、SHPR-NET[22]、CNN Model[23]、Pose REN[24]这些方法在精度上分别提高了 6.62%、0.48%、0.18%、4.31%、7.62%。在 ICVL 数据集上,所提算法较 Hand PointNet、3D DenseNet、SHPR-NET、CNN Model、Bayesian DeepPrior[25]、Pose REN、SO-HandNet、PCHPS[26]这些方法在精度上分别提高了 3.78%、1.43%、7.32%、6.01%、33.93%、1.72%、13.34%、24.96%。此外,虽然 PCHPS 方法在 MSRA 数据集上表现良好,但因该方法对不固定臂长、遮挡和自相似度过高的手势表现不佳,故其在 ICVL 数据集上的表现远远不如所提算法。所提算法在两个数据集上达到了最小的总体平均误差距离。

表 3　各方法在 MSRA 和 ICVL 数据集上的平均误差距离

Table 3　Average error distance of each method on MSRA and ICVL datasets　　unit: mm

Method	Mean error in MSRA	Mean error in ICVL
Hand PointNet[12]	8.505	6.935
3D DenseNet[21]	7.98	6.77
SHPR-NET[22]	7.96	7.2
CNN Model[23]	8.3	7.1
Bayesian DeepPrior[25]		10.1
Pose REN[24]	8.6	6.79
SO-HandNet[13]		7.7
PCHPS[26]	7.117	8.893
PointNet+MFSE	8.203	6.854
PointNet+FPR	8.192	6.728
PointNet+MFSE+FPR	7.942	6.673

3.5 实验结果定性展示

为了能够更直观地展示所提算法的效果，本节对所提算法进行了定性分析。图 6 和图 7 分别展示了在 ICVL 和 MSRA 数据集下，对于一些代表性的手势，所提算法的回归结果与真实值的对比。

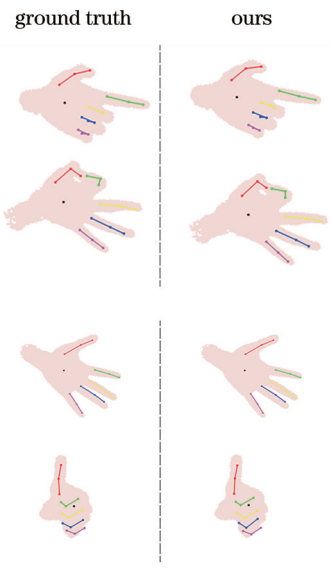

图 6 在 ICVL 数据集下的实验结果对比

Fig. 6 Comparison of experimental results under ICVL dataset

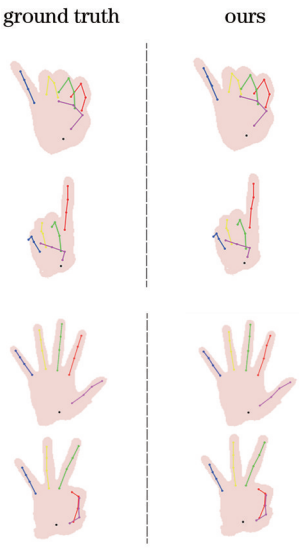

图 7 在 MSRA 数据集中的实验结果对比

Fig. 7 Comparison of experimental results under MSRA dataset

4 结 论

所提算法能提高点云网络对 3D 人手姿态估计的精度。FPR 策略作用于点云的预处理阶段，该策略通过置信度参数对手指位置处的点进行标记，从而强化对应点在网络特征提取阶段的作用，提高手指位置点特征的比例。MFSE 模块嵌入在网络的特征提取阶段，增强了分层网络中更有效的初始层局部特征的作用范围，并提高了这些有效特征对全局特征的贡献。在两个数据集上的实验都证明了所提方法达到了先进的 3D 人手姿态估计性能。

虽然所提算法能有效提高大部分手指关节点的回归精度，但是忽视了手指之间的差异性。未来的研究重点将在于提高 FPR 策略的普适性，研究更有效的特征融合方案。

参 考 文 献

[1] 徐志京，王东. 基于双路循环生成对抗网络的多姿态人脸识别方法[J]. 光学学报，2020，40(19)：1910002.
Xu Z J，Wang D. Multi-pose face recognition with two-cycle generative adversarial network[J]. Acta Optica Sinica，2020，40(19)：1910002.

[2] 程超，达飞鹏，王辰星，等. 基于 Lucas-Kanade 算法的最大 Gabor 相似度大姿态人脸识别[J]. 光学学报，2019，39(7)：0715005.
Cheng C，Da F P，Wang C X，et al. Pose invariant face recognition using maximum Gabor similarity based on Lucas-Kanade algorithm[J]. Acta Optica Sinica，2019，39(7)：0715005.

[3] 张政，徐杨. 基于双分类器的自适应单双手手势识别[J]. 激光与光电子学进展，2021，58(2)：0210005.
Zhang Z，Xu Y. Adaptive one-hand and two-hand gesture recognition based on double classifiers[J]. Laser & Optoelectronics Progress，2021，58(2)：0210005.

[4] 鲍志强，吕辰刚. 基于 Kinect 的实时手势识别[J]. 激光与光电子学进展，2018，55(3)：031008.
Bao Z Q，Lü C G. Real-time gesture recognition based on Kinect[J]. Laser & Optoelectronics Progress，2018，55(3)：031008.

[5] Mirsu R，Simion G，Caleanu C D，et al. A PointNet-based solution for 3D hand gesture recognition[J]. Sensors，2020，20(11)：3226.

[6] Oberweger M，Lepetit V. DeepPrior++：improving fast and accurate 3D hand pose estimation[C]//2017 IEEE International Conference on Computer Vision Workshops (ICCVW)，October 22-29，2017，Venice. New York：IEEE Press，2017：585-594.

[7] Xiong F，Zhang B S，Xiao Y，et al. A2J：anchor-to-joint regression network for 3D articulated pose estimation from a single depth image[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV)，October 27-November 2，2019，Seoul，Korea (South). New York：IEEE Press，2019：793-802.

[8] Malik J，Elhayek A，Stricker D. Structure-aware 3D hand pose regression from a single depth image[M]// Bourdot P，Cobb S，Interrante V，et al. Virtual reality and augmented reality. Lecture notes in computer science. Cham：Springer，2018，11162：3-17.

[9] Ge L H，Liang H，Yuan J S，et al. 3D convolutional neural networks for efficient and robust hand pose estimation from single depth images[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)，July 21-26，2017，Honolulu，HI. New York：

IEEE Press, 2017: 5679-5688.

[10] Chang J Y, Moon G, Lee K M. V2V-PoseNet: voxel-to-voxel prediction network for accurate 3D hand and human pose estimation from a single depth map[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 5079-5088.

[11] Malik J, Abdelaziz I, Elhayek A, et al. HandVoxNet: deep voxel-based network for 3D hand shape and pose estimation from a single depth map[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 7111-7120.

[12] Ge L H, Cai Y J, Weng J W, et al. Hand PointNet: 3D hand pose estimation using point sets[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 8417-8426.

[13] 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]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 6960-6969.

[14] Ge L H, Ren Z, Yuan J S. Point-to-point regression PointNet for 3D hand pose estimation[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11217: 489-505.

[15] Charles R Q, Hao S, Mo K C, et al. PointNet: deep learning on point sets for 3D classification and segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI. New York: IEEE Press, 2017: 77-85.

[16] Charles R Q, Yi L, Su H, et al. Pointnet++: deep hierarchical feature learning on point sets in a metric space [C]//The 24th Annual Conference on Neural Information Processing Systems, December 4-9, 2017, Long Beach, CA, USA. Cambridge: The MIT Press, 2017: 5105-5114.

[17] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT. New York: IEEE Press, 2018: 7132-7141.

[18] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[19] Sun X, Wei Y C, Liang S, et al. Cascaded hand pose regression[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 824-832.

[20] Tang D H, Chang H J, Tejani A, et al. Latent Regression Forest: structured estimation of 3D articulated hand posture[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE Press, 2014: 3786-3793.

[21] Ge L H, Liang H, Yuan J S, et al. Real-time 3D hand pose estimation with 3D convolutional neural networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(4): 956-970.

[22] Chen X H, Wang G J, Zhang C R, et al. SHPR-Net: deep semantic hand pose regression from point clouds[J]. IEEE Access, 2018, 6: 43425-43439.

[23] Ding L, Wang Y, Laganière R, et al. A CNN model for real time hand pose estimation[J]. Journal of Visual Communication and Image Representation, 2021, 79: 103200.

[24] Chen X H, Wang G J, Guo H K, et al. Pose guided structured region ensemble network for cascaded hand pose estimation[J]. Neurocomputing, 2020, 395: 138-149.

[25] Caramalau R, Bhattarai B, Kim T K. Active learning for Bayesian 3D hand pose estimation[C]//2021 IEEE Winter Conference on Applications of Computer Vision (WACV), January 3-8, 2021, Waikoloa, HI, USA. New York: IEEE Press, 2021: 3418-3427.

[26] Huang H Z, Zhuang Z L, Hu Q, et al. PCHPS: the estimation of 3D hand pose and shape using point cloud from a single depth image[C]//2020 IEEE 16th International Conference on Automation Science and Engineering (CASE), August 20-21, 2020, Hong Kong, China. New York: IEEE Press, 2020: 1231-1236.