doi：10. 12068/j. issn. 1005－3026. 2025. 20240079

基于几何注意力机制的三维手部姿势估计算法

邹 慧，佘黎煌，陈烨涵，乐 意
（东北大学 计算机科学与工程学院，辽宁 沈阳 110169）

摘 要：在Transformer编码-解码的基础架构上设计了手势识别网络，在自注意力机制的基础上引入了优化的偏移注意力机制来提取手部特征．同时为了更好地提取手部结构的局部特征，设计了邻域聚合策略．手部结构自身的三维复杂性导致其不同区域的平滑程度不同，进行手部姿势估计时，忽略这种特征会使手部结构的局部关键信息丢失，为了解决这一问题，对手部结构进行了几何分解，分别用锐变成分和柔变成分来表示手部结构的尖锐区域和平坦区域，通过注意力机制对这两种成分的特征给予不同的关注．在MSRA，ICVL和NYU数据集上的实验验证了此算法的准确度与SOTA算法相当．

关 键 词：手势识别；三维点云；注意力机制；Transformer模型；深度学习
中图分类号：TP 391.4 **文献标志码**：A **文章编号**：1005-3026(2025)10-0044-07

3D Gesture Estimation Algorithm Based on Geometric Attention Mechanism

ZOU Hui, *SHE Li-huang*, *CHEN Ye-han*, *YUE Yi*
(School of Computer Science & Engineering, Northeastern University, Shenyang 110169, China. Corresponding author: SHE Li-huang, E-mail: shelihuang@ise.neu.edu.cn)

Abstract: A gesture recognition network based on the coding and decoding infrastructure of Transformer was designed, and an optimized offset attention mechanism was introduced to extract hand features based on the self-attention mechanism. At the same time, in order to extract the local features of the hand structure better, a neighborhood aggregation strategy was designed. The three-dimensional (3D) complexity of the hand structure itself led to different levels of smoothness in different regions. When estimating gestures, ignoring this feature usually leads to the loss of local key information of the hand structure. In order to solve this problem, geometric decomposition of the hand structure was carried out, and sharp and flexible components were used to represent the sharp and flat regions of the hand structure, respectively. Different attention was paid to the characteristics of these two components through the attention mechanism. Experiments on MSRA, ICVL, and NYU datasets demonstrate that the accuracy of this algorithm is comparable to that of SOTA.

Key words: gesture recognition; 3D point cloud; attention mechanism; Transformer model; deep learning

手部姿势作为一种复杂的变形体，具有多样性、多义性以及时间上的差异等特征，这些特征使得手势识别具有挑战，现有的手势识别技术仍然存在着准确性与稳定性问题．手部姿势通常由手指和手腕的关节角度或旋转矩阵表示，通过计算关节点之间的相对位置和方向可以进行手势的估计．在手部姿势估计的相关文献和常用数据集中，常见的手部关节点个数为14，16，21，36.

现有的手部姿势估计的算法大多基于卷积神经网络（CNN）展开，通过CNN提取不同视角的图片特征，然后通过池化结合不同视角的信息．Ahmet等[1]使用三维卷积神经网络作为动态

收稿日期：2024-04-08
基金项目：辽宁省教育厅高等学校基本科研项目（LJKZ0011）；辽宁省科学技术计划项目（2021JH1/10400011）．
作者简介：邹 慧（2000—），女，吉林白山人，东北大学硕士研究生；佘黎煌（1980—），男，福建莆田人，东北大学讲师，博士．

手势检测器和分类器,通过分层架构实现实时动态手势识别.Chen 等[2]提出的 Pose-REN 基于单一的区域集成网络,直接回归手部关节位置.

V2V-PoseNet[3]使用 3D 体素化网格来估计 3D 关节点热图的算法,通过一个体素到体素的网络,实现在体素化深度图和体素化形状之间建立一对一的映射.体素化的算法能够将不规则的手部点云转化为规则的体素网格,但是体素网格的复杂度会随着体素分辨率的增加呈现指数级的增长.在体素化算法中,点云信息丢失的问题比较明显.直接处理点云的算法,则可以最大可能地保留原始信息,相关的研究有 Hand PointNet[4],So-HandNet[5]等.Hand PointNet 通过 PointNet++[6]提取手部特征,但是不规则的手部点云无法直接使用 CNN 处理,因此 Hand PointNet 需要首先对手部点云作标准化处理,然后将归一化的手部点云作为输入数据来回归手部姿势.

手部点云具有无序性、不规则性和稀疏性,Transformer 模型在处理连续的点时具有优秀的全局特征学习能力和置换等变的特性,适用于手部点云的处理与估计.Transformer 类模型在处理长序列和大规模数据时,可以捕捉序列中不同位置之间的依赖关系.本文采用 Transformer 编码-解码结构对输入的 3D 手部点云数据作输入嵌入处理.

手部的尖锐区域表示了手的轮廓特征,平坦区域表示了手的骨骼特征,受 GDANet[7]中点云分解策略的启发,本文设计了手部分解过程,即手部整体被分解为锐变分量和柔变分量,然后将其与原始的手部特征融合.将手部特征分解再融合的过程可以互补反映手的形状.三维手部姿势的几何分解与估计以手部点云的三维坐标作为输入,通过邻域嵌入和由数个堆叠的自注意力机制组成的偏移注意力机制提取手部特征,输出估计的手部关节位置信息.

1　算　　法

1.1　算法设计

本文设计了三维手部的几何分解过程,手部点云经过几何分解后再通过手部姿势回归网络提取三维手部特征.手部姿势回归网络由邻域嵌入模块和偏移注意力模块组成.邻域嵌入模块将手部点云映射到高维特征空间,将局部信息合并到嵌入特征中.这些特征在偏移注意力机制模块中进行特征提取.算法总体框图如图 1 所示.每个模块上方的数字表示输出通道,LBR 表示线性层、批归一化层和激活函数的结合,SG 为采样分组层,OA 为偏移注意力模块.

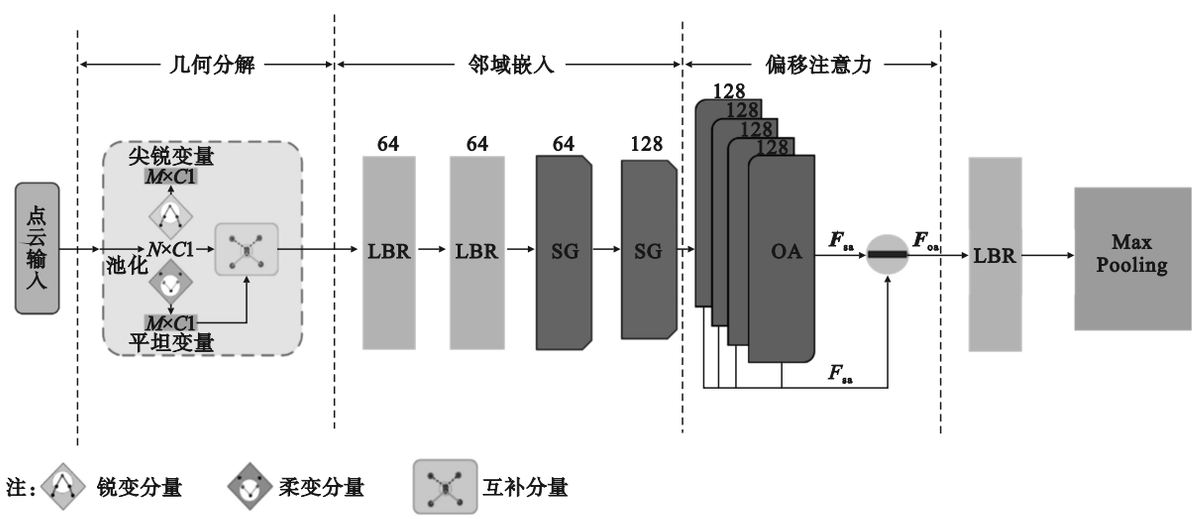

图 1　算法总体框图

Fig. 1　Overall block diagram of algorithm

1.2　三维手部几何分解

图卷积网络(GCN)可以捕获图像的全局和局部特征,但无法充分利用不同尺度之间的依赖关系,本文将三维手部点云分解成锐变分量和柔变分量,可以更好地理解点云信息.

本文将 N 个 C 维特征点组成的点云通过矩阵 $\boldsymbol{X}$ 表示:

$$\boldsymbol{X}=[x_1,x_2,\cdots,x_N]^{\mathrm{T}}=[s_1,s_2,\cdots,s_C]\in\mathbf{R}^{N\times C}.$$

其中:x_i 表示第 i 个点;s_C 表示第 C 个通道特征.通过邻接矩阵 $\boldsymbol{A}$ 对特征空间中点的相似性进行编码,构造一个图 $G=(v,\boldsymbol{A})$,每个点 x_i 与相应的图顶点 i 相关联,s_C 表示图信号,点 x_i 和 x_j 之间的边权重为

$$\boldsymbol{A}_{i,j}=\begin{cases} f(\left\| x_i-x_j \right\|_2), \left\| x_i-x_j \right\|_2<\tau; \\ 0, \text{其他}. \end{cases}$$

其中:f是一个非负递减函数,它保证矩阵$\boldsymbol{A}$是一个对角占优矩阵;τ表示阈值.为了处理不同点和特征尺度上大小的变化,将边权重归一化:

$$\bar{\boldsymbol{A}}_{i,j}=\frac{\boldsymbol{A}_{i,j}}{\sum_j \boldsymbol{A}_{i,j}}.$$

其中$\bar{\boldsymbol{A}}$是对角占优矩阵,获得一个图G,矩阵$\bar{\boldsymbol{A}}$的特征值从低到高表示图频率$(\tilde{\lambda}_1, \tilde{\lambda}_2, \cdots, \tilde{\lambda}_N)$.

在图像处理中,尖锐的边缘部分对应于高频响应,温和的平坦部分对应于低频响应.式(1)借助频率变化的原理,在图G上设计一个以图像信息作为输入的图形滤波器来选择属于三维手部的轮廓和平坦区域的点,可以实现将手部点云分解为尖锐和平坦的变化成分.

$$h(\boldsymbol{A})=\sum_{l=0}^{L-1} h_l \boldsymbol{A}^l. \tag{1}$$

其中:h_l是滤波器系数;L是滤波器长度.本文使用拉普拉斯算子,$L=2, h_0=1, h_1=-1$,图滤波器的多项式格式为$h(\bar{\boldsymbol{A}})=I-\bar{A}$.该滤波器以$s_C$作为输入,并生成一个过滤后的图信号$y_C$,由式(2)算出$h(\bar{\boldsymbol{A}})$的频率响应.

$$\widehat{h(\bar{\boldsymbol{A}})}=\begin{bmatrix} 1-\tilde{\lambda}_1 & 0 & \cdots & 0 \\ 0 & 1-\tilde{\lambda}_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1-\tilde{\lambda}_N \end{bmatrix}. \tag{2}$$

应用$h(\boldsymbol{A})$对点集进行过滤,得到一个过滤后的点集$h(\boldsymbol{A})\boldsymbol{X}$,由于$h(\bar{\boldsymbol{A}})=\boldsymbol{I}-\bar{\boldsymbol{A}}$,因此$h(\boldsymbol{A})\boldsymbol{X}$中的每个点可以表示为

$$(h(\bar{\boldsymbol{A}})\boldsymbol{X})_i=x_i-\sum_j^N \bar{\boldsymbol{A}}_{i,j}x_j.$$

当点x_i, x_j之间的距离小于阈值τ,$\boldsymbol{A}_{i,j}$保持非零值.$h(\bar{\boldsymbol{A}})\boldsymbol{X}$实际上等于一个点特征与其相邻特征的线性凸组合之间的差值,它反映了每个点对相邻点的变化程度.其中某些点的范数会发生明显变化,计算式(2)中每个点的$L2$范数,$L2$范数越大,说明变化越剧烈,也就意味着该点属于三维手部的轮廓,这与二维图像中通过高通滤波器获得边缘区域的原理一致.根据$L2$范数的值,所有的原始点降序排列为$\boldsymbol{X}_{\mathrm{o}}=[\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_N]^{\mathrm{T}}$,选择前$M$个点作为手部点云的锐变分量,最后$M$个点作为手部点云的柔变分量.

手部点云的原始特征是$\boldsymbol{X}_{\mathrm{o}}$,手部点云通过几何分解模块后,获得了锐变分量特征$\boldsymbol{X}_{\mathrm{s}}$,柔变分量特征$\boldsymbol{X}_{\mathrm{g}}$.利用这些特征进行不同的非线性编码,获得邻接矩阵:

$$\boldsymbol{W}_{\mathrm{s}}=\boldsymbol{\Theta}_{\mathrm{o}}(\boldsymbol{X}_{\mathrm{o}})\cdot\boldsymbol{\Theta}_{\mathrm{s}}(\boldsymbol{X}_{\mathrm{s}})^{\mathrm{T}},$$
$$\boldsymbol{W}_{\mathrm{g}}=\boldsymbol{\Phi}_{\mathrm{o}}(\boldsymbol{X}_{\mathrm{o}})\cdot\boldsymbol{\Phi}_{\mathrm{g}}(\boldsymbol{X}_{\mathrm{g}})^{\mathrm{T}}.$$

文中:$\boldsymbol{\Theta}, \boldsymbol{\Phi}, \boldsymbol{\Psi}$表示由不同的多层感知器实现的非线性函数.$\boldsymbol{W}_{\mathrm{s}}, \boldsymbol{W}_{\mathrm{g}}$的每一行分别与手部的锐变分量特征和柔变分量特征的注意力权重相关.如图2所示将两种分量的特征与手部点云的原始特征进行互补融合,并对它们给予不同的关注,最后这些特征通过式(3)被连接为互补分量.

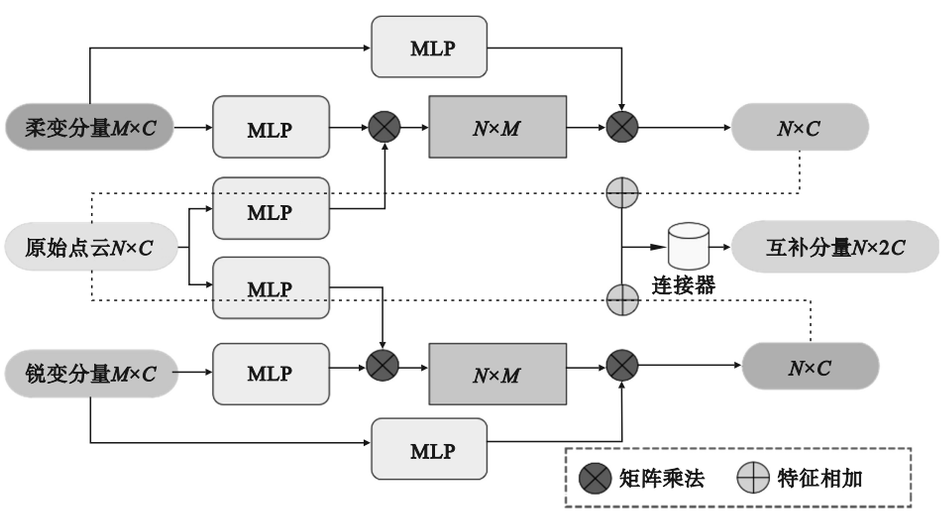

图2 手部几何分解示意图

Fig. 2 Schematic diagram of geometric decomposition of hand

$$\left.\begin{aligned} \boldsymbol{Y}_{\mathrm{s}}&=\boldsymbol{X}_{\mathrm{o}}+\boldsymbol{W}_{\mathrm{s}}\cdot\boldsymbol{\Psi}_{\mathrm{s}}(\boldsymbol{X}_{\mathrm{s}}), \\ \boldsymbol{Y}_{\mathrm{g}}&=\boldsymbol{X}_{\mathrm{o}}+\boldsymbol{W}_{\mathrm{g}}\cdot\boldsymbol{\Psi}_{\mathrm{g}}(\boldsymbol{X}_{\mathrm{g}}), \\ \boldsymbol{Z}&=\boldsymbol{Y}_{\mathrm{s}}\oplus\boldsymbol{Y}_{\mathrm{g}}. \end{aligned}\right\} \tag{3}$$

将手部点云的锐变分量和柔变分量视作两种表征,这两种表征在三维手部的整体重构过程中起到互补的作用.

几何分解模块将手部点云分解为包含M个点的锐变分量和包含M个点的柔变分量.通过分析手部点云的几何结构和特征,将点云中的关键点与非关键点进行区分,使得网络更加注重关键点的重要信息,提高关键点的利用率.

1.3　手部姿势回归网络

本文提出的三维手部姿势估计算法将包含手部关节点的深度图像作为输入,并输出一组三维手部关节位置$\Phi=\{\phi_t\}_{t=1}^{T}\in\Lambda$,其中$T$是手部关节数量.手部点云经过三维手部几何分解模块后,通过邻域嵌入模块被编码到一个新的高维特征空间.4个堆叠的注意力模块为每个手部点学习一个语义丰富且具有辨别力的表征,然后由一个线性层来生成输出特征.

在训练阶段给定N个具有人工标注的手部关节位置的训练样本$\{X_i,\Phi_i\}_i^N$,对训练集中的手部关节位置进行主成分分析,可以获得手部关节位置的投影α_i,最小化目标函数获得使误差最小的最佳网络参数:

$$\omega^*=\arg\min\sum_{i=1}^{N}\|\alpha_i-P(X_i,\omega)\|^2+\lambda\|\omega\|^2.$$

其中:ω为训练获得的网络参数.

手部识别网络输出的手部关节信息:

$$\hat{\Phi}=E\cdot P(\boldsymbol{X},\omega^*)+u.$$

其中:E为主成分;P为手势识别网络;$\boldsymbol{X}$为网络的输入;u为经验均值.

本文使用编码-解码结构来提取手部点云的特征,编码部分主要包括1个邻域嵌入模块和4个堆叠的注意模块.邻域嵌入模块由2个线性层和2个SG(采样和分组)层组成,用以转化手部点云的坐标,将三维的手部点云坐标映射到一个更高维的空间,使具有相似语义信息的手部点云能够在高维空间中更加靠近.

手部点云经过两级SG层时,点云大小分别减小到512和256.两个级联的SG层在特征聚合过程中逐渐扩大感受野,在点云取样过程中使用欧氏距离对每个点进行KNN搜索分组,从本地聚集特征.例如,第2次SG的过程如图3所示,将512个点和相应的特征$\boldsymbol{F}$作为输入,输出256个采样点和聚合特征$\boldsymbol{F}_s$.

经过转化后的坐标输入到级联的自注意力网络中,学习三维手部结构的点特征.每个注意层的输出维度与输入维度相同,将各个层级的自注意力网络的输出拼接起来,获得自注意力特征$\boldsymbol{F}_{sa}$,再通过元素减法计算自注意力特征$\boldsymbol{F}_{sa}$与输入特征$\boldsymbol{F}_{in}$之间的偏移量,即偏移注意力$\boldsymbol{F}_{oa}$.

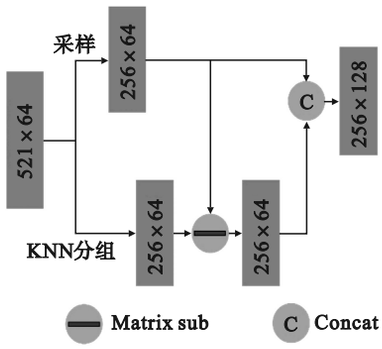

图3　SG采样分组示意图

Fig. 3　Schematic diagram of SG sampling grouping

1.4　邻域嵌入

Transformer模型是纯粹的注意力模型,它不能自然地处理输入序列之间的相对位置关系,需要使用位置编码器来显式地编码这种位置信息.而手部点云本身含有三维坐标,因此在模型的输入部分,本文将手部点云的坐标进行了输入嵌入,可以增强模型提取手部局部特征的能力.

首先对输入的手部点云$\boldsymbol{X}\in\mathbf{R}^N$和相应的特征$\boldsymbol{F}$进行采样和分组,并且通过最近邻算法(KNN)找出每个采样点的k邻域,计算每个域中的点与采样点之间的差值,然后根据式(4)将这个差值与采样点连接起来.

$$\left.\begin{aligned}\Delta\boldsymbol{F}(x)&=\mathrm{concat}(\boldsymbol{F}(q)-\boldsymbol{F}(p))\overrightarrow{AB},\\ \tilde{\boldsymbol{F}}(x)&=\mathrm{concat}(\Delta\boldsymbol{F}(x),\mathrm{RP}(\boldsymbol{F}(x),k)),\\ \boldsymbol{F}_s(x)&=\max-\mathrm{pooling}(L(L(\tilde{\boldsymbol{F}}(x)))).\end{aligned}\right\}\tag{4}$$

其中RP($\boldsymbol{F}(x)$,k)是向量$\boldsymbol{F}(x)$重复k次形成矩阵的算子.最终通过两层全连接层和最大池化得到每个采样区域的局部特征.

1.5　偏移注意力机制

在图卷积网络(GCN)中用拉普拉斯矩阵来替代邻接矩阵,可以聚合节点附近的邻接信息并把节点自身的信息也考虑进去.文献[8]中提出,在对点云应用Transformer时,用拉普拉斯矩阵$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{E}$来替代对角线度矩阵$\boldsymbol{E}$,可以获得更好的网络性能,本文据此嵌入了偏移注意模块.

将手部的自注意力特征$\boldsymbol{F}_{sa}$与输入特征$\boldsymbol{F}_{in}$之间的差值作为输出特征,可以有效提升网络的性能.输入特征的线性变换产生$\boldsymbol{Q}$(查询矩阵)、$\boldsymbol{K}$(键矩阵)、$\boldsymbol{V}$(值矩阵).

$(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\boldsymbol{F}_{in}\cdot(\boldsymbol{W}_q,\boldsymbol{W}_k,\boldsymbol{W}_v)$.其中$\boldsymbol{W}_q,\boldsymbol{W}_k,\boldsymbol{W}_v$是共享的可学习的线性变换.然后通过式(5)的矩阵点积来计算注意权重:

$$\tilde{\boldsymbol{A}}=(\tilde{\alpha})_{i,j}=\boldsymbol{Q}\times\boldsymbol{K}^{\mathrm{T}}.\tag{5}$$

自注意输出特征 $\boldsymbol{F}_{sa}=\boldsymbol{AV}$ 是使用相应的注意权值的值向量的加权和.

输出特征 $\boldsymbol{F}_{oa}=\mathrm{OA}(\boldsymbol{F}_{in})=\mathrm{LBR}(\boldsymbol{F}_{sa}-\boldsymbol{F}_{in})+\boldsymbol{F}_{in}$, $\boldsymbol{F}_{sa}-\boldsymbol{F}_{in}$ 相当于离散拉普拉斯算子,通过式(6)计算得到.

$$\boldsymbol{F}_{in}-\boldsymbol{F}_{sa}=\boldsymbol{F}_{in}-\boldsymbol{AV}=\boldsymbol{F}_{in}-\boldsymbol{AF}_{in}\boldsymbol{W}_v\approx \boldsymbol{F}_{in}-\boldsymbol{AF}_{in}=(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{F}_{in}\approx L\boldsymbol{F}_{in}. \tag{6}$$

其中: $\boldsymbol{I}$ 是单位矩阵,类似于拉普拉斯矩阵的对角度矩阵 $\boldsymbol{D}$; $\boldsymbol{A}$ 是注意矩阵,类似于邻接矩阵 $\boldsymbol{E}$; $\boldsymbol{W}_v$ 是LBR层的权重矩阵,在计算过程中被忽略.

2 实 验

2.1 实验设置

本文的模型在MSRA[9], NYU[10], ICVL[11]手势数据集上做了评估实验.MSRA数据集包含76 k深度图像.包含9个由17个手势组成的主题,每个图像帧有21个手部关节,包括每根手指的4个关节和手掌的1个关节.用8个样本集训练,用剩余的样本集进行测试.NYU数据集由72 k训练深度图像和8.2 k测试深度图像组成,手势标记包含36个关节,实验使用14个关节点的子集,即2个手腕关节点、1个手掌关节点和11个手指关节点.ICVL数据集由330 k训练深度图像和1.6 k测试深度图像组成,手势标记包含16个关节,包括15个手指关节点和1个手掌关节点.网络输出的手部关节信息为原手部关节信息维数的三分之二.

2.2 对比分析

本实验将估计的三维手部关节点的位置和手部骨骼形态显示在深度图像上,然后与数据集上的真实标记对比.在MSRA, NYU, ICVL数据集上的定性结果如图4所示,图中点为关节点,实线表示骨骼形态,第1行为估计的结果,第2行为数据集上的真实标记,可见本文算法可以准确估计出手部姿势.

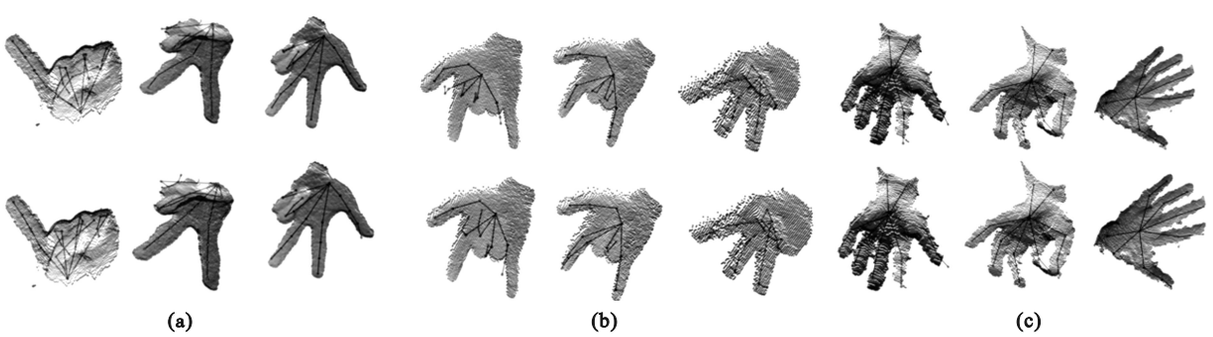

图4 手势估计结果示意图

Fig. 4 Schematic diagram of gesture estimation results

(a)—MSRA; (b)—NYU; (c)—ICVL.

各种算法的平均误差距离的对比见表1.本文算法在MSRA数据集上的平均误差比Hand Pointnet低0.4 mm,在NYU数据集上的平均误差比DeepPrior++[12]低0.21 mm,在ICVL数据集上的平均误差比Hand Pointnet低0.2 mm.可见本文算法在平均误差距离指标上优于其他算法.各关节点的平均误差距离的对比见图5,可见本文算法有良好的表现.模型在MSRA数据集上的质量分析见图6.本文对比了各算法的高质量估算结果占比,在相同误差范围内,高质量估算结果占比越高,表明模型质量越好.通过比较可知本模型能够完成高质量的手势识别.

表1 平均误差距离的对比

Table 1 Comparison of mean error distance

方法	平均误差/mm NYU	方法	平均误差/mm ICVL	方法	平均误差/mm MSRA
3DCNN[13]	14.1	DeepMode1[17]	11.6	Cascade	15.2
Pose-REN	11.8	Cascade	9.9	Occlusion[19]	12.8
DeepPrior++	12.2	CrossingNets[18]	10.2	CrossingNets	12.2
REN-9×6×6[14]	12.7	DeepPrior++	8.1	REN-9×6×6	9.7
Feedback[15]	16.0	LRF	12.6	DeepPrior++	9.5
Hand3D[16]	17.6	Hand PointNet	6.9	Hand PointNet	8.5
本文	12.0	本文	6.7	本文	8.1

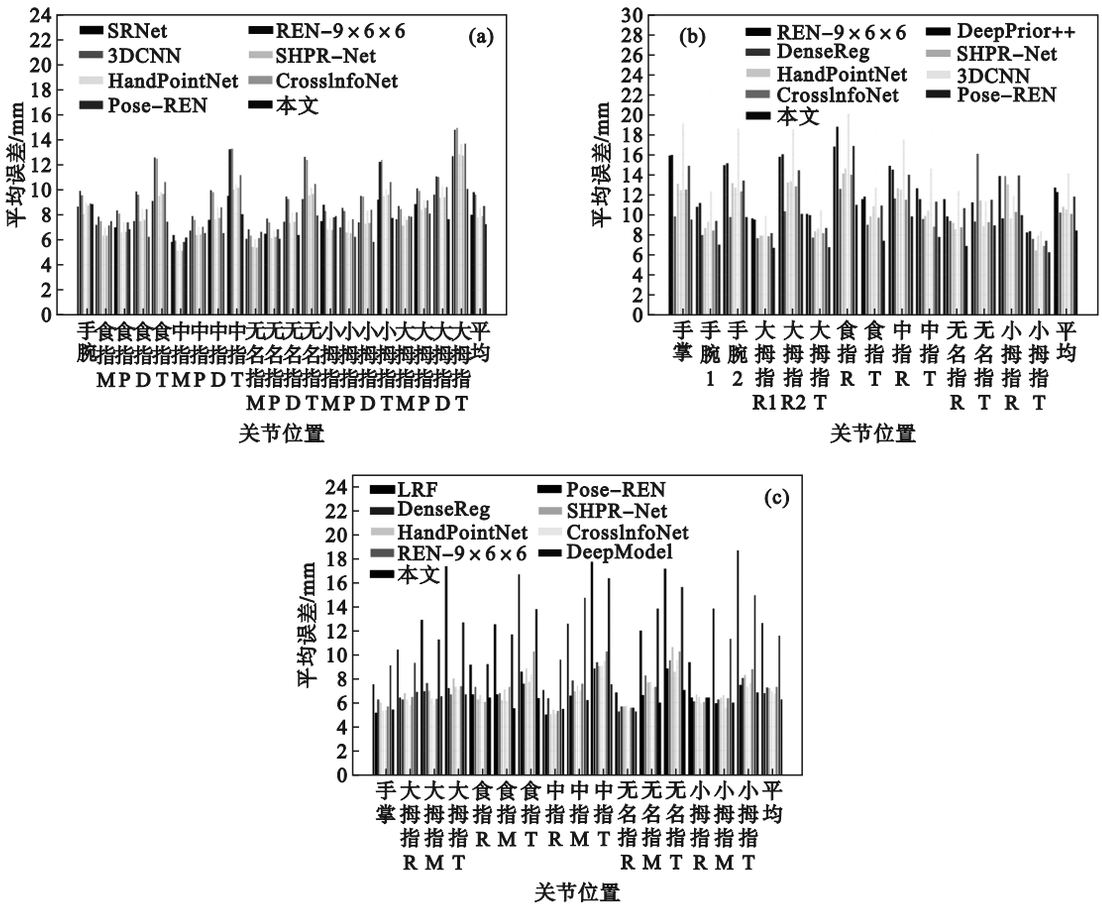

图5　各关节点的平均误差距离的对比

Fig. 5　Comparison of average error distance of each node

(a)—MSRA; (b)—NYU; (c)—ICVL.

2.3　消融实验

消融实验结果见表2. 当输入手部点云与锐变和柔变成分的特征相融合时,网络实现的误差为8.1 mm是最好的效果,而手部点云只与自身融合时效果最差,证明了本文给予手部点云的锐变分量与柔变分量不同的关注有助于更好理解手部结构的局部特征.

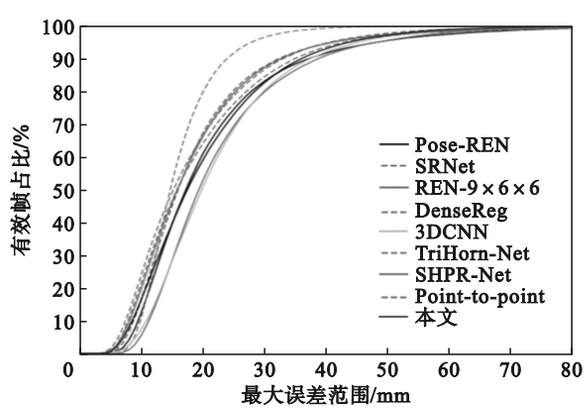

图6　模型质量分析

Fig. 6　Model quality analysis

2.4　稳定性分析

通过1 024个点作为网络的输入来测试模型的稳健性.由图7可以看出,均方误差曲线不断下降并趋于稳定,训练集与测试集对于该网络结构是收敛稳定的,证明了基于几何注意力机制的三维手部姿势估计算法是稳定的.

表2　互补注意效果

Table 2　Complementary attention effect

自身融合	锐变分量	柔变分量	误差/mm
√			8.6
	√		8.3
		√	8.4
	√	√	8.1

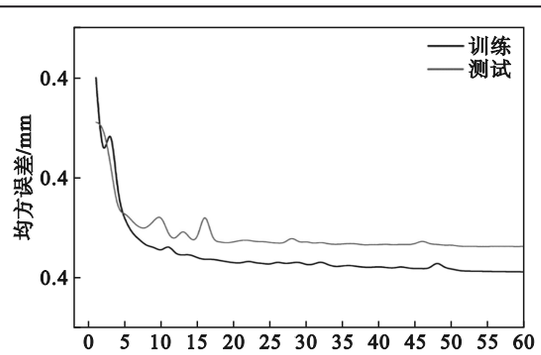

图7　均方误差

Fig. 7　Mean squared error

3 结 语

本文提出了一种新的手部姿势估计算法,基于Transformer架构,引入了能够聚合邻域信息的输入嵌入模块,在对手部点云进行坐标嵌入的同时提取手部结构的局部特征.计算自注意力特征和输入特征之间的差值获得偏移注意力机制,能够提高模型的性能.将手部结构进行几何分解,对锐变分量和柔变分量给予不同的关注,从几何角度补充对手部点云的理解.本文算法在公开手部数据集上进行的实验取得了较优的表现,可用于三维手部姿势的估计.

参考文献:

[1] Ahmet G, Neslihan K, Gerhard R, et al. Real-time hand gesture detection and classification using convolutional neural networks [C]// 2019 IEEE International Conference on Automatic Face and Gesture Recognition. Lille, 2019: 1-8.

[2] 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.

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

[4] Ge L H, Cai Y J, Weng J W, et al. Hand PointNet: 3D hand pose estimation using point set [C]// 2018 IEEE International Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 8417-8426.

[5] 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 International Conference on Computer Vision. Seoul, 2019: 6960-6969.

[6] Qi C R, Yi L, Su H, et al. PointNet++: deep hierarchical feature learning on point sets in a metric space [J]. *Advances in Neural Information Processing Systems*, 2017, 30(10): 5105-5114.

[7] Xu M T, Zhang J H, Zhou Z P. Learning geometry-disentangled representation for complementary understanding of 3D object point cloud [C]// 2021 AAAI Conference on Artificial Intelligence. Vancouver, 2021, 35(4): 3056-3064.

[8] Guo M H, Liu Z J. Point cloud transformer [J]. *Computational Visual Media*, 2021, 7(2): 187-199.

[9] Sun X, Wei Y C, Liang S, et al. Cascaded hand pose regression [C]// 2015 IEEE International Conference on Computer Vision and Pattern Recognition. Boston, 2015: 824-832.

[10] 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.

[11] Tang D H, Chang C J, Alykhan, et al. Latent regression forest: structured estimation of 3D articulated hand posture [C]// 2014 IEEE International Conference on Computer Vision and Pattern Recognition. Columbus, 2014: 3786-3793.

[12] Markus O, Vincent L. Deepprior++: improving fast and accurate 3D hand pose estimation. [C]// 2017 IEEE International Conference on Computer Vision Workshop. Venice, 2017: 585-594.

[13] 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 International Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 5679-5688.

[14] Wang G J, Chen X H, Guo H K, et al. Region ensemble network: towards good practices for deep 3D hand pose estimation [J]. *Journal of Visual Communication and Image Representation*, 2018, 55(8): 404-414.

[15] Markus O, Wohlhart P, Lepetit V. Training a feedback loop for hand pose estimation [C]// 2015 IEEE/CVF International Conference on Computer Vision. Santiago, 2015: 3316-3324.

[16] Deng X M, Yang S, Zhang Y D, et al. Hand3D: hand pose estimation using 3D neural network [C]// 2017 IEEE/CVF International Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 549-557.

[17] Zhou X Y, Wan Q F, Zhang W, et al. Model-based deep hand pose estimation [C]// 2016 International Joint Conference on Artificial Intelligence. New York, 2016: 2421-2427.

[18] Wan C D, Probst T, Luc V G, et al. Crossing nets: combining GANs and VAEs with a shared latent space for hand pose estimation [C]// 2017 IEEE/CVF International Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 1196-1205.

[19] Madadi M, Escalera S, Carruesco A, et al. Occlusion aware hand pose recovery from ` sequences of depth images [C]// 2017 International Conference on Automatic Face & Gesture Recognition. Washington DC, 2017: 230-237.