doi:10.3969/j.issn.1001-4616.2025.06.010

基于注意力卷积混合架构的实时手部网格重建

杜逸飞,李　琳,孔京辉

(合肥工业大学计算机与信息学院(人工智能学院),安徽 合肥 230601)

[**摘要**]　当前的手部网格重建方法主要关注手部网格重建的精度和偏差,而对于其实时性方面的因素关注则较少. 为此,本文提出了一种基于三阶段注意力卷积混合架构的三维手部网格重建方法(QuickHand). 首先,设计了一个混合层级特征提取器,通过将轻量级卷积编码、多尺度特征变换和全局注意力机制有机结合,实现了从单视图手部图像到判别性特征表示的高效映射. 其次,构建了一个维度映射转换器,通过自适应的关节特征编码和空间变换,实现了从二维平面特征到网格顶点特征空间的精确转换. 最后,设计了一个高效的网格重建解码器,通过深度可分离螺旋卷积和多级上采样策略,在保持低计算复杂度的同时实现了高精度的手部三维网格重建. 实验表明,本文在保持高精度手部网格重建的同时提高了实时性能,相比面向精度的方法具有更高的实时性,相比面向轻量化的方法具有更优的重建精度.

[**关键词**]　三维手部网格重建,混合层级特征提取,实时性,维度映射转换,网格重建解码

[**中图分类号**]TP391　[**文献标志码**]A　[**文章编号**]1001-4616(2025)06-0090-11

Real-Time Hand Mesh Reconstruction Based on Attention-Convolution Hybrid Architecture

Du Yifei, Li Lin, Kong Jinghui

(Visualization and Cooperative Computing(VCC) Laboratory, School of Computer Science and Information Engineering(School of Artificial Intelligence), Hefei University of Technology, Hefei 230601, China)

Abstract: Current hand mesh reconstruction methods primarily focus on the accuracy and deviation of hand mesh reconstruction, with less attention paid to the real-time performance of hand mesh reconstruction. To address this, this paper proposes QuickHand, a three-dimensional hand mesh reconstruction method based on a three-stage attention-convolution hybrid architecture. Firstly, a hybrid hierarchical feature extractor is designed, which achieves efficient mapping from single-view hand images to discriminative feature representations by combining lightweight convolutional encoding, multi-scale feature transformation, and a global attention mechanism. Secondly, a dimension mapping transformer is constructed, which enables precise conversion from 2D planar features to mesh vertex feature spaces through adaptive joint feature encoding and spatial transformation. Finally, an efficient mesh reconstruction decoder is designed, which achieves high-precision 3D hand mesh reconstruction while maintaining low computational complexity through depth-separable spiral convolution and multi-level upsampling strategies. Experiments demonstrate that the proposed method achieves real-time performance while maintaining high-precision hand mesh reconstruction, offering better real-time performance compared to accuracy-oriented methods and superior reconstruction accuracy compared to lightweight-oriented methods.

Key words: 3D hand mesh reconstruction, hybrid hierarchical feature extraction, real-time performance, dimension mapping transformation, mesh reconstruction decoding

三维手部网格重建技术经历了从传统方法到深度学习驱动的革命性转变. 早期研究主要依赖于传统计算机视觉技术,如随机森林分类器、高斯混合模型和基于模板的匹配方法[1-3]. 这些传统方法通常需要精心设计的手工特征提取过程,且在复杂背景、光照变化和严重遮挡情况下表现不佳,泛化能力有限. 随

收稿日期:2025-03-23.

基金项目:国家自然科学基金面上项目(62277014).

通讯作者:李琳,博士,副教授,研究方向:虚拟现实与人机交互. E-mail:lilin_julia@hfut.edu.cn

着深度学习的蓬勃发展,手部重建研究逐渐形成了两条平行但相互影响的技术路线:一条专注于提升重建精度,另一条则致力于在保证足够精度的前提下提升推理速度,以满足实际应用场景的需求.

深度卷积神经网络(CNN)凭借其强大的特征提取和表示学习能力,迅速成为手部网格重建的主流方法. 早期 CNN 模型[4-6]主要基于合成数据或有限的真实数据进行训练,后续随着研究的深入,高质量标注数据集如 FreiHAND[7]、HO3D[8]、InterHand2.6M[9]和 DexYCB[10]逐渐建立. 这些数据集是通过精心设计的采集系统获取的,FreiHAND 和 HO3D 数据集提供了高质量的 RGB 图像以及对应的三维手部姿态和形状标注,InterHand2.6M 数据集提供了大规模的双手交互场景,而 DexYCB 数据集则专注于手部抓取各种日常物体的视频序列,这些数据集共同为研究提供了宝贵的训练资源. 研究人员基于这些数据集,设计了从回归关键点到直接预测网格顶点的各类网络架构[11-16],推动了重建精度的不断提升. 然而,这类监督学习方法在应对复杂背景、极端姿态和严重遮挡等挑战性场景时,仍存在一定的精度下降和泛化困难. 同时,现有的 CNN 架构往往计算量大、参数众多,难以满足实时交互应用的速度需求.

近年来,Transformer 架构凭借其强大的长距离依赖建模能力,为手部重建领域注入了新的活力. 基于 Transformer 的方法[17-19]能够更好地捕捉手部关节之间的空间关系和解剖学约束,显著提升了复杂姿态下的重建精度. 一些研究将 Transformer 与图神经网络结合[20-21],利用手部骨骼结构的先验知识来引导特征学习和信息传递. 尽管如此,Transformer 模型通常参数量庞大、计算开销高,难以在资源受限设备上实时运行;同时,现有 Transformer 方法主要关注静态单帧重建,对动态手部运动序列的建模不足,限制了其在交互场景中的应用效果. 然而,现有的手部 Transformer 方法存在泛化能力有限的问题,且仅有少数研究系统性地探讨了实时性手部网格重建的问题. 在实际应用环境中,手部经常与物体或其他身体部位发生复杂交互和遮挡,现有方法在这些场景下的表现仍有较大提升空间.

在实时性研究方向上,网络轻量化成为关键挑战. 研究者通过模型压缩、知识蒸馏、网络剪枝和量化等技术,大幅减少了模型参数量和计算复杂度. 一些工作[16,22]引入了计算高效的网络组件,如深度可分离卷积、递归架构等,在维持重建质量的同时提升了推理速度. 然而,这些轻量化方法往往以牺牲重建精度为代价,特别是在复杂姿态和细节表现上存在明显退化;同时,大多数实时方法仅考虑单帧重建,缺乏对时序一致性的建模,在实际应用中影响了用户体验和系统的准确性.

为应对现实场景中的复杂挑战,研究领域拓展出多个分支. 多模态融合方法结合深度信息[23]、热图[24-25]和二维关键点[13,26]等多种输入,显著增强了系统在复杂背景和遮挡情况下的鲁棒性. 参数化手部模型如 MANO 的引入[16,20,23,25,27]使网络能够学习低维度的形状和姿态参数,而非直接预测高维网格顶点,有效降低了学习难度并提高了重建质量. 此外,弱监督和自监督学习策略[6,15,26]的发展减轻了对大规模三维标注数据的依赖,通过利用二维投影一致性、时序连贯性和对称性等先验知识来指导网络学习. 尽管如此,多模态融合方法通常需要专门的传感设备,限制了其在普通消费设备上的应用;参数化模型虽然降低了学习复杂度,但对于非标准手部形状的表达能力有限;弱监督方法则在极端姿态下的精度仍有较大提升空间.

手部网格重建任务根据交互对象和场景复杂度可分为几个主要类别:单手重建[13,16-18,26-27]聚焦于单只手的精确建模,主要挑战在于捕捉复杂的指间关节运动和手指微小变形;而双手交互重建[12,28]则需处理两只手之间的复杂接触、遮挡和相互作用关系,这类任务要求模型理解手与手之间的物理约束和空间关系. 手部与物体交互重建任务[2-3]中,需同时建模手部形变和物体位置,特别是在接触区域的精确重建尤为困难,模型需要理解手部与不同物体之间的交互方式和接触动力学. 当手部被遮挡为主要任务[21,24]时,要求模型能够从有限可见区域推断完整手部结构,通常依赖时序信息或先验知识来补全被遮挡部分,在现实应用场景中具有重要意义. 然而,这些任务在现实场景中仍面临诸多挑战,包括跨数据集泛化能力不足、实时性与精度之间的权衡困难、极端姿态下的重建不稳定以及大规模高质量标注数据获取成本高昂等问题,亟待更先进的算法和更高效的数据利用策略来解决. 针对上述问题,本文提出了一个实时性的手部网格重建管道 QuickHand,该系统经过优化设计,能够在 NVIDIA GeForce RTX 3060 Ti GPU 上实现 95FPS 的处理速度,可以使得手部跟踪和重建流畅运行. 该管线包含一种混合层级的特征提取器和一种碰撞损失函数. 混合层级的特征提取器通过将轻量级卷积编码、多尺度特征变换和全局注意力机制有机结合,使网络能够同时捕捉手部的局部细节特征和全局上下文信息;碰撞损失函数可以防止手指间非自然穿透,实现高精度、实时的手部模型重建.

1　研究方法

本文参考 Mobrecon[16] 网络架构,提出了一种三维手部网格重建方法 QuickHand,主要是基于 Transformer 和 CNN 结合实现三维手部网格重建. 考虑到准确重建手部网格需要充分捕捉顶点间的空间关联性,本文设计了混合层级特征提取器,通过引入 Transformer 增强了对全局特征的提取能力. 同时,为了防止生成的手部网格出现非物理穿透,使生成的手部更自然,引入了碰撞损失来减少手部间的碰撞. QuickHand 目标是生成预测的顶点信息 P 与面片信息 S,其表示公式如下:

$$P=\{p_i\}_{i=1},S=\{s_i\}_{i=1}. \tag{1}$$

其整体网络架构如图 1 所示,首先,输入的 RGB 图片通过特征提取模块的混合层级特征提取器进行高效的特征提取,该模块融合了卷积神经网络的局部感受野和注意力机制的全局上下文捕获能力. 随后,通过维度映射转换器实现了二维到三维的特征映射,该模块利用坐标映射技术将平面特征投影至三维空间. 最后,利用深度可分离螺旋卷积对三维网格进行高效解码,该方法通过分离通道维度和空间维度的卷积运算,在保持高质量几何重建能力的同时,显著降低了计算复杂度.

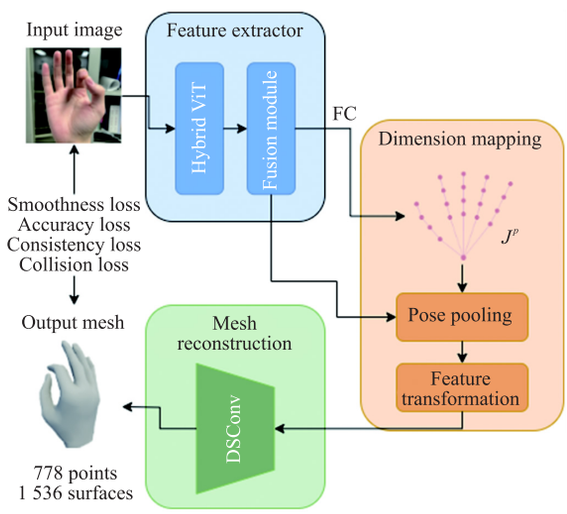

图 1　QuickHand 网络架构

Fig. 1　Network framework of QuickHand

1.1　混合层级特征提取器

特征提取对整个手部模型重建过程的速度有很大影响,因此使用占用内存空间小、运行速度快的特征提取方法将会有效提升手部模型重建过程. 为了实现高效而全面的特征提取,本文采用了混合视觉变换器架构,如图 2 所示.

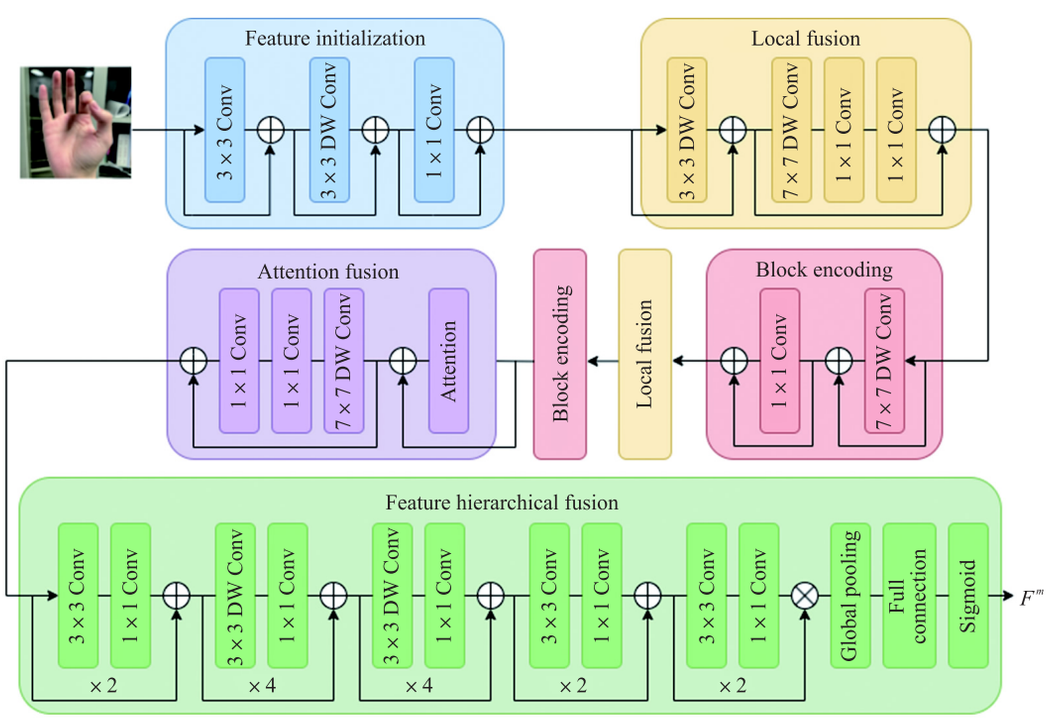

图 2　混合层级特征提取器架构

Fig. 2　Framework of hybrid hierarchical feature extractor

该模块由四个部分构成:

(1)特征初始化层,负责初步下采样和通道调整.

(2)分块编码层,负责进行高效特征分块.

(3)局部融合层,通过局部特征混合来进行特征变换:

$$\boldsymbol{O}=DWConvBNI, \tag{2}$$

其中 $\boldsymbol{O}$ 为输出张量,I 为输入张量,$DWConvBNI$ 表示将 I 批量归一化后进行深度可分离卷积操作.

(4)注意力融合层,加入 Attention 机制,在保持计算效率的前提下有效捕获全局上下文信息.

这种层次化设计使网络同时具备 CNN 的局部感受野和 Transformer 的长距离依赖建模能力. 技术亮点在于结构重参数化策略——训练时保持多分支架构提升学习能力,推理时将多分支融合为单一卷积. 具体实现方法是在训练阶段采用包含批归一化、深度卷积和跳跃连接的多分支结构来增强特征表达能力,而在推理阶段将整个多分支架构重参数化为单一的深度卷积层,从而显著降低内存访问成本和推理延迟. 此外,混合视觉变换器通过深度可分离卷积,在各阶段构建了丰富的多尺度特征表示,显著提升了计算效率与特征表达能力的平衡性.

特征层级融合器对特征进行融合. 基于混合视觉变换器提取的特征,需要进一步增强其表达能力并适应三维重建任务. 为此,本文设计了特征层级融合器对其进行有效传递与融合:

$$\boldsymbol{F}^m \in \mathbf{R}^{S^m\times H^m\times W^m}, \tag{3}$$

其中,$\boldsymbol{F}^m$ 表示潜在特征编码,S^m、H^m、W^m 是张量通道的大小(Size)、高度(Height)和宽度(Width),m 表示融合器输出的特征,$\mathbf{R}$ 是实数集,表示张量中所有元素均为实数.

特征层级融合器由多个特征融合单元串联组成,每个特征融合单元内部通过稠密连接方式将各层特征充分融合,并在每个单元之后自适应地调整各通道特征的重要性权重. 这种结构设计具有显著优势:一方面促进了特征的多尺度整合,增强了模型对手部细节的表达能力;另一方面改善了梯度流动,提升了训练稳定性,同时增强了特征表示的判别性. 在整个特征融合过程中,通过多层次的上采样和跳跃连接操作,逐步将深层语义特征与浅层空间细节信息融合,最终生成维度适配的特征表示,为后续的二维到三维映射提供了理想的特征基础.

1.2 维度映射转换器

在本节中将详细介绍提出的维度映射转换器,该模块旨在实现从二维特征表示到三维重建所需信息的转换. 该模块将混合层级特征提取器输出的特征通过潜在特征编码和二维关节点预测,最终转化为适用于三维重建的顶点特征表示.

(1)特征降维与潜在特征编码. 从混合层级特征提取器获得的特征包含了丰富的手部信息,但维度较高导致计算量大,因此需要去除冗余信息,映射为一种更适合三维重建任务的特征形式. 降维不仅减少了后续处理的计算量,更重要的是使网络学习更为精炼的特征表示,提取对手部重建最为关键的几何和姿态信息.

(2)二维关节点坐标预测. 精确的关节点定位对于手部重建至关重要. 在本文的实验中,首先进行通道降维,然后重塑为适合关节点表示的形式,最后通过全连接网络直接预测 21 个手部关节点的二维坐标. 与基于热图的间接方法相比[24-25],这种直接回归方式具有更高的计算效率和潜在的精度优势,特别是在处理密集排列的手指关节时表现更为稳定.

(3)姿态池化. 完成二维关节点坐标预测后,此时获得了手部关节的空间位置,但仍需要提取这些位置对应的手部姿态特征. 因此,本文使用直接回归方法预测二维位置时,通过网格采样技术(一种在特征图上预定位置采样特征值以保留空间结构信息的方法)在潜在特征编码中进行姿态池化,公式如下:

$$\boldsymbol{F}^p=\{\boldsymbol{F}^m(J_i)\}_{i=1,2,\cdots,N}, \tag{4}$$

其中,J 为预测的二维关节点坐标,N 为手部关节点数量. 这一过程将特征转换为位姿对齐特征 $\boldsymbol{F}^p\in\mathbf{R}^{N\times C^m}$,建立了每个关节点与特征空间的精确对应关系. 其中 C^m 是 $\boldsymbol{F}^m$ 的通道数.

(4)特征空间变换. 为了将关节点特征映射到更密集的顶点特征空间,本文设计了特征空间变换机制. 由于 MANO 风格的手部网格包含 778 个顶点和 21 个关节点,直接映射存在挑战. 首先将模板网格下采样至最小尺寸(49 个顶点),然后通过转换矩阵进行映射:

$$\boldsymbol{F}^{\min p}=\boldsymbol{M}\times\boldsymbol{F}^{p}, \tag{5}$$

其中,$\boldsymbol{M}\in\mathbf{R}^{V^{\min}\times N}$为可学习的转换矩阵,$V^{\min}$表示最小尺寸网格的顶点数,$\boldsymbol{F}^{\min p}$为顶点特征表示. 其中可学习的转换矩阵采用小数值 0.01 均匀分布的方式进行初始化,并在训练过程中,通过反向传播算法进行优化,根据损失函数学习从 2D 关键点特征空间到 3D 顶点空间的最优映射关系. 这种线性转换操作降低了计算复杂度,提升了模型效率.

通过上述多步转换的设计,维度映射转换器建立了从二维特征表示到三维顶点特征的有效桥梁,在降低计算复杂度的同时保证了重建的准确性. 实验表明,该方法在重建质量和计算效率方面均取得了显著提升.

1.3　网格重建解码器

在本节中将详细介绍提出的网格重建解码器,该部分旨在将网格顶点特征重建为完整的手部三维网格. 受到以往工作中螺旋卷积在处理网格数据上的优秀表现启发,本文使用了一种基于深度可分离设计的高效螺旋卷积,同时设计了相应的解码器结构,在保持重建精度的同时显著降低了计算开销.

深度可分离螺旋卷积. 在处理网格数据时,普通的螺旋卷积虽然能有效利用顶点的邻域信息,但计算复杂度较高. 为此,本文使用深度可分离螺旋卷积,将传统的螺旋卷积分解为深度卷积和点卷积两个步骤:

$$\boldsymbol{F}_{p}^{d}=\left\{W_{i}^{d}\times\boldsymbol{F}_{p',i}\right\}_{i=1}^{\boldsymbol{D}}, \tag{6}$$

$$\boldsymbol{F}_{p}^{p}=W^{p}\times\boldsymbol{F}_{p}^{d}, \tag{7}$$

其中,d 表示深度卷积 DSConv,上角标 p 表示点卷积,下角标 p 和 p'表示点,$\boldsymbol{D}$ 表示特征向量的维度 dimension. $p'\in n\text{-}sp(p)_{l}$,$n\text{-}sp(p)_{l}$ 表示顶点 p 的螺旋邻域,$\boldsymbol{F}_{p}^{d}$ 表示顶点 p 经过深度卷积操作后的特征表示,$\boldsymbol{F}_{p',i}$表示顶点 p'的第 i 维特征,$\boldsymbol{F}_{p}^{p}$ 表示顶点 p 经过点卷积操作后的特征表示,W^{d} 和 W^{p} 分别为深度卷积和点卷积的学习参数.

三维解码器. 为了实现高质量的网格重建,本文设计了一个三维解码器结构. 包含上采样、深度可分离螺旋卷积和非线性激活操作. 解码器接收最小尺寸的网格特征(49 个顶点),通过多次上采样和特征转换,逐步恢复到完整的手部网格(778 个顶点).

1.4　损失函数设计

为了提升手部重建的准确性和时序一致性,本文设计了多任务联合损失函数:

重建精度损失. 首先,使用 $L1$ 损失监督二维关节点和三维网格顶点的预测:

$$L_{J2D}=\|J-J^{g}\|, \tag{8}$$

$$L_{M}=\|p-p^{g}\|, \tag{9}$$

其中,p 和 J 分别表示预测的三维网格顶点和二维关节点坐标,g 表示对应的 Groundtruth,$J2D$ 表示二维关节点,M 表示三维网格顶点. 这两项损失确保了重建结果在三维和二维空间的几何准确性.

网格平滑度损失. 为了保证重建网格的表面平滑性,本文引入法向量损失和边缘损失:

$$L_{N}=\sum_{s\in S}\sum_{(i,j)\in s}\left|\frac{p_{i}-p_{j}}{\|p_{i}-p_{j}\|}\times\boldsymbol{n}_{s}\right|, \tag{10}$$

$$L_{E}=\sum_{s\in S}\sum_{(i,j)\in s}\left|\|p_{i}-p_{j}\|-\|p_{i}^{g}-p_{j}^{g}\|\right|, \tag{11}$$

其中,S 为网格面片集合,$\boldsymbol{n}_{s}$ 为面片 s 的单位法向量,$\boldsymbol{N}$ 表示法向量 normal,E 表示边缘 edge. 法向量损失通过约束相邻顶点的方向关系来保持局部平滑性,边缘损失则确保网格边的长度与模板一致.

碰撞损失. 为了避免手指之间的非物理穿透,本文引入碰撞损失:

$$L_{C}=\sum_{s\in S}\sum_{(i,j)\in s}\max(0,d_{0}-\|p_{i}-p_{j}\|), \tag{12}$$

其中,d_{0} 为预定义的最小距离阈值,C 表示碰撞 collide,该损失函数惩罚任何距离小于 d_{0} 的顶点对,从而防止网格的自交叉,使重建结果更符合物理约束.

一致性损失. 为了增强模型对视角变化的鲁棒性,本文基于数据增强引入二维和三维的一致性损失. 具体而言,对输入图像进行仿射变换和颜色增强,生成两个增强视图:

$$V=\left\{v_{i}\right\}_{i=1,2}, \tag{13}$$

$$L_{C3D} = \| P_{v1}^{Rot} - P_{v2}^{Rot} \|, \tag{14}$$

$$L_{C2D} = \| J_{v1}^{Tran} - J_{v2}^{Tran} \|, \tag{15}$$

其中,*Tran* 和 *Rot* 分别表示两个视图间的相对仿射变换和相对旋转变换,*V* 表示视图 view,*C3D* 表示三维一致性损失,*C2D* 表示二维一致性损失. 这两项损失确保了在不同视角和时间点的重建结果保持一致性.

完整的损失函数为:

$$L_{\text{TOTAL}} = L_{J2D} + L_M + \lambda L_N + L_E + \lambda L_C + L_{C3D} + L_{C2D}, \tag{16}$$

其中,$\lambda = 0.1$,即法向量损失和碰撞损失权重设置为 0.1,其余损失函数权重均为 1.

实验表明,这种多任务损失函数的设计不仅保证了几何准确性和表面平滑性,避免了非物理的网格穿透现象,还确保了在不同视角和时间点的重建结果保持一致性.

2 实验设计

2.1 实验环境及参数设置

实验在一台配置为 Intel i5-13600、16GB 内存、Nvidia RTX3060 Ti 显卡的机器上进行模型的训练和评估. 模型基于 PyTorch 搭建,使用 Adam 优化器,batch_size 设置为 32,epoch 设置为 60,初始学习率设置为 10^{-3},在第 30 个 epoch 时将学习率降低至原来的 0.1 倍. 超参数设置中,输入分辨率为 512×512,螺旋邻域大小 l 设置为 9,阈值 d_0 设置为 10 mm. 模型采用端到端可训练的方式进行训练,通过最小化损失函数来促使模型更有效地学习手部网格的三维表示.

2.2 数据集及评估指标

实验采用了两个主流的手部重建数据集进行评估:FreiHAND 和 HO3Dv2. FreiHAND 包含 130 240 张训练图像和 3 960 张测试图像,每张图像都标注了手部的三维网格和二维关节点. HO3Dv2 是一个手-物交互数据集,包含 66 034 张训练图像和 11 524 张测试图像,提供了具有挑战性的物体遮挡场景.

为了全面评估模型性能,本文采用以下评价指标:

PA-MPVPE(Procrustes Analysis Mean Per-Vertex Position Error):经过对齐后的平均每个顶点的位置误差,消除了全局姿态差异的影响,用于评估三维网格重建的精度,后续实验中简称 PV,其计算公式如下:

$$\text{PA-MPVPE} = \frac{1}{N}\sum_{i=1}^{N} \| p_i - p_i^g \|, \tag{17}$$

PA-MPJPE(Procrustes Analysis Mean Per-Joint Position Error):经过对齐后的平均每个关节点的位置误差,消除了全局姿态差异的影响,用于评估三维关节点预测的精度,后续实验中简称 PJ,其计算公式如下:

$$\text{PA-MPJPE} = \frac{1}{N}\sum_{i=1}^{N} \| J_i - J_i^g \|, \tag{18}$$

F-Score:网格重建质量的评估指标,@5 mm 和@15 mm 分别表示距离阈值为 5 mm 和 15 mm 时的 *F* 值,其计算公式如下:

$$\text{Precision} = \frac{TP}{TP+FP}, \tag{19}$$

$$\text{Recall} = \frac{TP}{TP+FN}, \tag{20}$$

$$\text{F-score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}, \tag{21}$$

其中,Precision 表示精确率;Recall 表示召回率;*TP* 表示正确预测的正例,即真阳;*FP* 表示错误预测的正例,即假阳;*FN* 表示错误预测的反例,即假阴.

FPS(Frames Per Second):每秒处理的图像帧数,用于评估模型的实时性能,其计算公式如下:

$$\text{FPS} = \frac{NIP}{T}, \tag{22}$$

其中,*NIP* 为处理的图像总帧数,*T* 为处理总时间.

3　实验结果

3.1　对比试验

在本节中,将所提出的方法与目前主流的方法进行了比较,这些方法包括 SMHR[29]、CMR[30]、MobRecon[16]和 HaMer[31]. 实验在 FreiHAND 和 HO3Dv2 数据集上评估了方法的性能,其中 FreiHAND 数据集结果如表 1 所示,HO3Dv2 数据集结果如表 2 所示. 本文使用 224×224 分辨率的输入图片进行验证,对比结果如图 3 所示. 同时,为了验证所提出方法的泛化性,本文使用 512×512 分辨率的真实拍摄图片进行验证,如表 3 和图 4 所示.

表 1　FreiHAND 数据集对比实验

Table 1　Comparative experiments on FreiHAND dataset

方法	PV/mm	PJ/mm	*F*@5	*F*@15	FPS/(img/s)
SMHR	8.0	7.8	0.649	0.966	15
CMR	7.5	7.4	0.680	0.973	53
MobRecon	7.2	6.9	0.694	0.979	67
HaMer	5.7	6.0	0.785	0.990	8
Ours	6.4	6.3	0.747	0.984	92

表 2　HO3D 数据集对比实验

Table 2　Comparative experiments on HO3D dataset

方法	PV/mm	PJ/mm	*F*@5	*F*@15	FPS/(img/s)
SMHR	12.5	12.1	0.502	0.951	15
CMR	11.2	10.9	0.519	0.952	53
MobRecon	9.4	9.2	0.538	0.957	67
HaMer	7.9	7.7	0.635	0.980	8
Ours	8.8	8.6	0.618	0.969	92

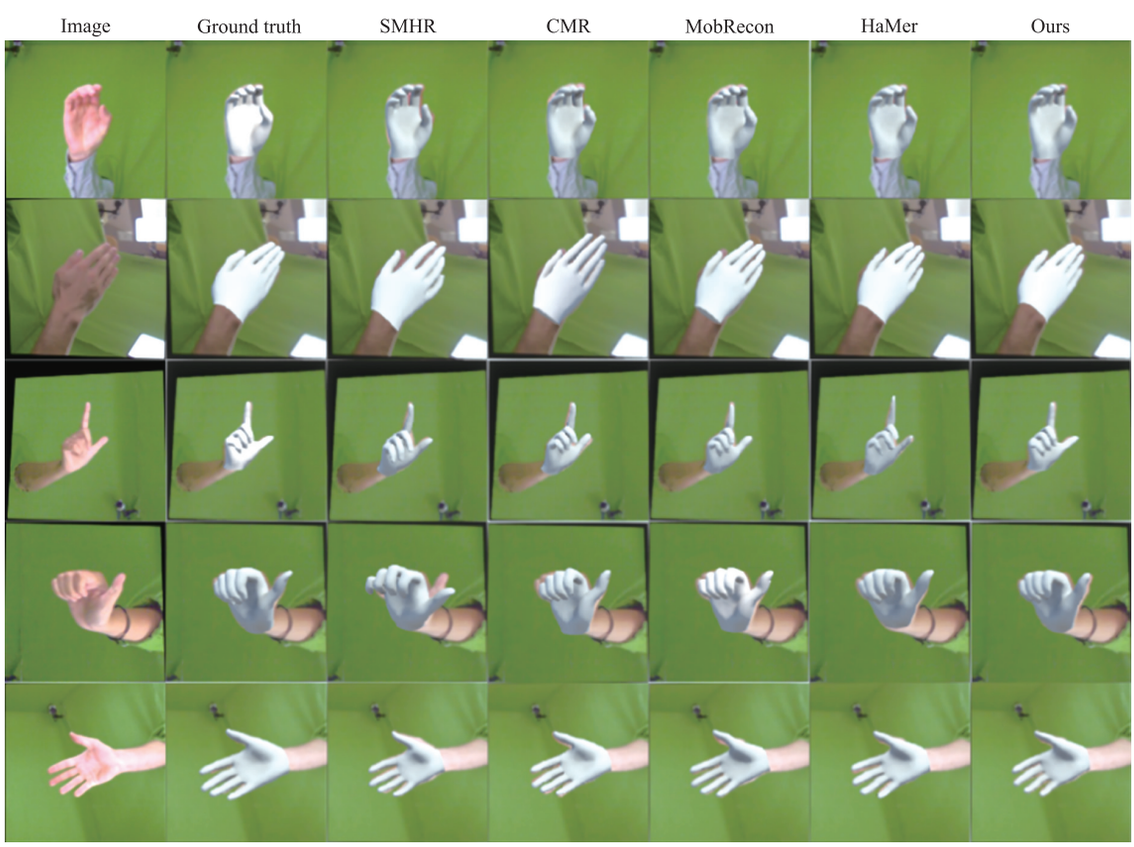

图 3　对比试验结果图

Fig. 3　Comparative experiments results

表 3　真实图片定量结果

Table 3　Real image quantitative results

方法	PV/mm	PJ/mm	*F*@5	*F*@15	FPS/(img/s)
SMHR	14.4	14.1	0.429	0.899	9
CMR	12.2	12.0	0.508	0.953	46
MobRecon	10.5	10.1	0.521	0.955	55
HaMer	9.0	8.8	0.567	0.963	5
Ours	10.3	9.8	0.525	0.954	78

图 4　泛化性验证图

Fig. 4　Generalization test results

在 FreiHAND 和 HO3D 数据集上进行的对比实验结果表明,在精度方面,本文的方法高于 SMHR、CMR 和 Mobrecon,略低于 HaMer;实时性方面,本文的方法高于其余 4 种方法,保证了手部网格重建的实时性. 本文的方法在确保手部网格重建质量没有明显下降的同时,达到了实时性的目的.

在泛化性评估方面,本文的方法有更强的跨场景适应能力. 当面对实际应用环境中真实拍摄的图片时,本文的方法在手部网格重建精度和实时性方面表现更好,泛化效果更显著. 这一特性有利于网络的实际部署,确保了系统在复杂多变的现实环境中的可用性和鲁棒性.

3.2　消融实验

3.2.1　对混合层级特征提取器层级配置进行讨论

为了探究混合层级特征提取器中卷积层和自注意力层数量配置对模型性能的影响,本文进行了详细的消融实验. 具体而言,本文设计了 5 种不同的网络配置,包括 3 层和 4 层两种深度,并在其中探索了自注意力层的不同配置方案.

首先,本文评估了 3 层结构的两种配置,如表 4 所示. 3Conv 和 2Conv-1Att 分别代表全卷积结构和引入一层自注意力的结构. 结果显示,2Conv-1Att 在保持 FPS 指标的同时精度有了显著提升,这说明适当引入自注意力机制能在保证推理速度的同时提高生成网格精度. 这是因为自注意力层能够高效地捕获特征图中的全局关联性,加强了手部不同区域特征之间的信息交互.

表 4　3 层配置结果

Table 4　Three-layer configuration results

模型	PV/mm	PJ/mm	Params/M	FPS/(img/s)
3Conv	7.2	6.9	5M	96
2Conv-1Att	6.4	6.3	5M	92

接下来,本文探究了 4 层结构的 3 种不同配置,如表 5 所示. 4Conv 代表纯卷积结构,3Conv-1Att 和 2Conv-2Att 分别代表引入一层和两层自注意力的结构. 其中,3Conv-1Att 取得了精度和速度的较好平衡,但相比 3 层结构的 2Conv-1Att,额外层并未带来显著的性能提升,且速度出现明显下降.

综合所有配置的实验结果,2Conv-1Att 结构在保持较高重建精度的同时,实现了最优的推理速度. 这表明,在本文的任务中,适度的网络深度配合恰当的自注意力机制是最有效的设计选择. 这种配置既保证了足够的特征提取能力,又避免了冗余计算,为实际应用提供了最佳的性能权衡方案. 本文使用 512×512 分辨率的输入图像进行效果验证,如图 5 所示.

表 5　4 层配置结果

Table 5　Four-layer configuration results

模型	PV/mm	PJ/mm	Params/M	FPS/(img/s)
4Conv	7	6.9	8.2M	68
3Conv-1Att	6.3	6.2	8.2M	62
2Conv-2Att	6.5	6.2	8.2M	59

图 5　不同配置效果验证

Fig. 5　Effectiveness verification of different configurations

3.2.2　对碰撞损失函数进行讨论

为了验证碰撞损失函数的有效性,本文对碰撞损失函数进行了消融实验. 通过对比有无碰撞损失函

数的重建结果,评估其对模型性能的影响.

首先,本文评估了移除碰撞损失函数的影响. 训练结果如表 6 所示,QuickHand 和 QuickHand-LColl 分别代表加入和移除碰撞损失前后的模型. 结果显示,移除碰撞损失后,模型在 PV 和 PJ 指标上出现了上升,这表明碰撞约束提高了重建精度. 相比之下,加入碰撞损失的模型提高了重建精度. 这表明在三维手部重建任务中引入碰撞检测是必要的,能够有效提升重建结果的准确性.

表 6　碰撞损失函数消融结果

Table 6　Ablation results of collision loss function

模型	PV/mm	PJ/mm	*F*@5	*F*@15
QuickHand	6.4	6.3	0.747	0.984
QuickHand-LColl	6.6	6.5	0.731	0.977

4　结论

本文提出了一种基于三阶段注意力卷积混合架构的三维手部网格重建方法 QuickHand. 该方法首先通过混合层级特征提取器实现了高效的特征提取,然后利用维度映射转换器将平面特征转换到三维网格顶点空间,最后采用轻量级解码器生成高精度手部网格. 实验表明,QuickHand 在手部网格重建任务上实现了精度和实时性的有效平衡. 同时,该方法仍存在一些局限性:对非标准手部形状(如残疾或特殊手势)的适应性有待提高;在极端视角遮挡或复杂背景下重建效果仍有提升空间,在多样化移动设备上的部署还需进一步探索. 未来研究可探索更具适应性的参数化手部模型以增强对特殊手形的重建能力,引入时序一致性约束提升动态序列重建质量,以及通过多模态融合进一步提高重建精度.

[参考文献]

[1] KESKIN C,KIRAÇ F,KARA Y E,et al. Hand pose estimation and hand shape classification using multi-layered randomized decision forests [C]//Computer Vision-ECCV 2012: 12th European Conference on Computer Vision. UK: Springer International Publishing,2012:852-863.

[2] TANG D,JIN CHANG 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. USA:CVPR,2014:3786-3793.

[3] ROMERO J,KJELLSTRÖM H,KRAGIC D. Hands in action:real-time 3D reconstruction of hands in interaction with objects [C]//2010 IEEE International Conference on Robotics and Automation. USA:ICACC,2010:458-463.

[4] MUELLER F,BERNARD F,SOTNYCHENKO O,et al. Ganerated hands for real-time 3d hand tracking from monocular rgb [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA:CVPR,2018:49-59.

[5] ZHANG X,LI Q,MO H,et al. End-to-end hand mesh recovery from a monocular rgb image[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. USA:CVF,2019:2354-2364.

[6] MOON G,SHIRATORI T,LEE K M. Deephandmesh:a weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling[C]//Computer Vision-ECCV 2020:16th European Conference. UK:Springer International Publishing,2020:440-455.

[7] ZIMMERMANN C,CEYLAN D,YANG J,et al. Freihand:a dataset for markerless capture of hand pose and shape from single rgb images[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. USA:CVF,2019:813-822.

[8] HAMPALI S,RAD M,OBERWEGER M,et al. Honnotate:a method for 3d annotation of hand and object poses[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. USA:CVPR,2020:3196-3206.

[9] MOON G,YU S I,WEN H,et al. Interhand2.6m:A dataset and baseline for 3d interacting hand pose estimation from a single rgb image[C]//Computer Vision-ECCV 2020:16th European Conference. UK:Springer International Publishing,2020:548-564.

[10] CHAO Y W,YANG W,XIANG Y,et al. DexYCB:a benchmark for capturing hand grasping of objects[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. USA:CVPR,2021:9044-9053.

[11] YANG L,XU J,ZHONG L,et al. POEM:reconstructing hand in a point embedded multi-view stereo[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. USA:CVPR,2023:21108-21117.

[12] CHEN X, SONG Z, JIANG X, et al. HandOS: 3D hand reconstruction in one stage[J/OL]. arXiv Preprint arXiv: 2412.01537, 2024.

[13] CHEN X, LIU Y, MA C, et al. Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. USA: CVF, 2021: 13274-13283.

[14] LUAN T, ZHAI Y, MENG J, et al. High fidelity 3d hand shape reconstruction via scalable graph frequency decomposition [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. USA: CVPR, 2023: 16795-16804.

[15] ZHENG X, WEN C, XUE Z, et al. HaMuCo: Hand pose estimation via multiview collaborative self-supervised learning[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. USA: CVF, 2023: 20763-20773.

[16] CHEN X, LIU Y, DONG Y, et al. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. USA: CVPR, 2022: 20544-20554.

[17] HUANG L, TAN J, LIU J, et al. Hand-transformer: Non-autoregressive structured modeling for 3d hand pose estimation[C]// Computer Vision-ECCV 2020: 16th European Conference. UK: Springer International Publishing, 2020: 17-33.

[18] CHENG W, KIM E, KO J H. HandDAGT: A denoising adaptive graph transformer for 3D hand pose estimation[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 35-52.

[19] FU Q, LIU X, XU R, et al. Deformer: Dynamic fusion transformer for robust hand pose estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. USA: CVF, 2023: 23600-23611.

[20] KONG D, ZHANG L, CHEN L, et al. Identity-aware hand mesh estimation and personalization from rgb images[C]// European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 536-553.

[21] PARK J K, OH Y, MOON G, et al. Handoccnet: Occlusion-robust 3d hand mesh estimation network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. USA: CVF, 2022: 1496-1505.

[22] CHENG W, KO J H. Handr2n2: Iterative 3d hand pose estimation using a residual recurrent neural network[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. USA: CVF, 2023: 20904-20913.

[23] FANG L, LIU X, LIU L, et al. Jgr-p2o: Joint graph reasoning based pixel-to-offset prediction network for 3d hand pose estimation from a single depth image [C]//Computer Vision-ECCV 2020: 16th European Conference. UK: Springer International Publishing, 2020: 120-137.

[24] IQBAL U, MOLCHANOV P, GALL T B J, et al. Hand pose estimation *via* latent 2.5d heatmap regression[C]//Proceedings of the European Conference on Computer Vision(ECCV). Italy: ECCV, 2018: 118-134.

[25] 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]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. USA: CVF, 2020: 7113-7122.

[26] KULON D, GULER R A, KOKKINOS I, et al. Weakly-supervised mesh-convolutional hand reconstruction in the wild[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. USA: CVPR, 2020: 4990-5000.

[27] MOON G, LEE K M. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image[C]//Computer Vision-ECCV 2020: 16th European Conference. UK: Springer International Publishing, 2020: 752-768.

[28] POTAMIAS R A, ZHANG J, DENG J, et al. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild[J/OL]. arXiv Preprint arXiv: 2409.12259, 2024.

[29] REN J, ZHU J, ZHANG J. End-to-end weakly-supervised single-stage multiple 3D hand mesh reconstruction from a single RGB image[J]. Computer vision and image understanding, 2023, 232: 103706.

[30] CHEN X, LIU Y, MA C, et al. Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. USA: CVPR, 2021: 13274-13283.

[31] PAVLAKOS G, SHAN D, RADOSAVOVIC I, et al. Reconstructing hands in 3d with transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. USA: CVF, 2024: 9826-9836.

[责任编辑: 黄 敏]