文章编号:1671-6833(2026)05-0009-08

基于图卷积网络的三维手部姿态估计

彭春燕[1,2], 王 璇[1,2], 陈杨博[1,2], 何港波[1,2]

(1. 青海师范大学 计算机学院,青海 西宁 810016;2. 青海师范大学 藏语智能全国重点实验室,青海 西宁 810016)

摘 要: 基于单张彩色图片的三维手部姿态估计由于手部存在遮挡、手部自相似性高等原因使预测结果存在误差大、手部结构不自然等问题。针对这些问题,首先,提出一个基于图卷积的三维手部姿态估计方法,使用 Keypoint R-CNN 提取图像视觉特征和手部关键点二维位置信息,将特征信息输入到改进的自适应核图卷积模块(AK_GraFormer)中;其次,引入带残差连接的 AKGNN 图核,自适应处理图数据以增强模型的特征学习与表达;最后,利用提出的评估指标监控动态训练策略以获得更优的估计结果。在 HO-3D_v3 数据集与 FreiHand 数据集上进行实验,结果表明:在单张彩色图片手部三维姿态估计任务中,所提方法相比其他同类方法具有明显优势,刚性对齐后的平均每关节位置误差(*PA-MPJPE*)最高降低了 12.50 百分点,检测关节点百分比曲线下面积(*AUC*)最高提高了 3.44 百分点。

关键词:三维手部姿态估计; 图卷积网络; 特征提取; 图核学习优化; 评估指标动态调整

中图分类号:TP391;TP751 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2026.02.013

手是人们与世界感知、交互的重要媒介,是连接人类与世界的桥梁。计算机通常需要检测手部的姿态特征来理解人类行为,因此,获取手-物图片并从中估计手部关键点和物体的三维位置极其重要。另外,虚拟现实技术蓬勃发展,人机交互使其应用领域加以拓展,如何通过获取三维手部姿态以精准操作虚拟物体成为人机交互的研究热点问题[1]。

近年来,大量的研究聚焦于独立的三维手部姿态估计上,而手-物三维姿态估计相较于单独的人手估计存在更多遮挡,因此更具挑战性。手部关节多,自由度高,形状变换相比人体更为复杂,手部关节相似性高,因而对于复杂手势识别较为困难。另外,手部关节灵活性高,物体具有一定体量,手-物图片中往往存在不同程度的自遮挡与相互遮挡,导致手部三维姿态估计准确度低且误差较大。

目前,手部姿态估计方法按照手部姿态的生成方式可分为以下 3 类:生成式手部姿态估计方法、判别式手部姿态估计方法和混合式手部姿态估计方法。生成式手部姿态估计方法通过使用预先定义的手部三维模型、复杂的目标函数与优化求解算法来进行姿态估计。Oikonomidis 等[2]将手部参数所定义的手的三维几何模型进行渲染得到合成图像,通过约束合成图像的手部轮廓与真实图像的手部轮廓之间的误差能量函数,结合粒子群优化算法求解模型参数。研究者分别提出了 sphere-meshes 模型[3]、MANO 手模型[4]、SMPL-X 手模型[5],随后大量研究基于这些模型展开。判别式手部姿态估计方法直接从输入的手部数据中提取具有判别力的特征,学习手部特征与其姿态空间的映射。早期的判别式手部姿态估计方法大都采用机器学习的方法,如随机森林算法[6]。随着深度学习的发展,出现了越来越多的基于深度学习的方法。Tompson 等[7]是将卷积神经网络(convolutional neural networks, CNN)应用到手部姿态估计的先驱者,随后手部姿态估计领域涌现出诸多基于卷积神经网络的手部姿态估计方法[8-10]。混合式手部姿态估计方法则是将生成式和判别式方法组合到一个框架中,一般先使用判别式方法的结果作为生成式方法的初始化[11],更容易优化求解得到更好的结果,但是相比前两种生成方式,混合式模型结构相对复杂,且计算成本较高。

收稿日期:2026-04-04;**修订日期**:2026-06-27

基金项目:国家自然科学基金资助项目(62441609,62563033);青海省实验室建设项目(2025-ZJ-J08)

作者简介:彭春燕(1980—),女,山东菏泽人,青海师范大学教授,博士,主要从事文化计算、机器学习的研究,E-mail:pcy@qhnu.edu.cn。

引用本文:彭春燕,王璇,陈杨博,等. 基于图卷积网络的三维手部姿态估计[J]. 郑州大学学报(工学版),2026,47(5):9-16. [Peng Chunyan, Wang Xuan, Chen Yangbo, et al. 3D hand pose estimation based on graph convolution network [J]. Journal of Zhengzhou University(Engineering Science), 2026,47(5):9-16.]

近期,基于图卷积网络(graph convolutional networks, GCN)从二维姿态估计三维姿态的方法,已取得非常优异的成果[12-14]。由于使用单个二维关键点来估计三维中的对应点是一个不确定性问题,CNN 的全连接层只能回归它们的位置,无法对关键点之间的连接进行建模,而 GCN 是可以显式处理手关节链图结构的专用网络,能够利用学习手部结构回归节点三维位置。Doosti 等[12]提出了 HOPE-Net 模型,将手部骨骼的拓扑结构看作图结构,通过学习节点关系得到手部姿态。Zhao 等[13]提出的 GraFormer 模型,通过结合多头自注意力机制和图卷积操作,在三维姿态估计任务中取得了较好的效果。Cai 等[15]通过在关节与相邻帧中的对应关节之间创建额外的边缘,从几个时间相邻的二维身体姿势中创建了一个时空图。Aboukhadra 等[16]将 GCNs 和多头注意力层进行组合,改善了手-物体交互方面的有效性。Zhuang 等[17]基于运动学的抓取稳定性,利用手和物体的初始姿态,将二维坐标与提取的特征拼接成图形。Zhang 等[18]提出了一种手图和对象图之间的交互感知方法。马胜蕾等[19]提出基于双分支多尺度注意力的手部三维姿态估计,对手关节之间的复杂关联关系建立模型。Yang 等[20]提出一种集多头自注意力、基于空间的图卷积和基于频谱的图卷积于一体的三维手部姿态和网格估计框架。

上述方法利用图卷积对手部三维姿态估计的算法均提出了创新性方法,然而在估计的准确度以及模型的表达能力方面仍有进一步提升的空间。本文基于图卷积模型,使用 Keypoint R-CNN[21],从单张 RGB 图像中提取多个二维特征,如热图、边界框、特征图等,经过变换得到二维位置信息和图像视觉特征,然后将这些特征信息输入到 AK_GraFormer 模块中。该模块利用 GraFormer[13]和 AKGNN[22],通过使用改进的 res_AKGNN 层使图卷积网络进一步学习特征表示,进而提升模型的表达能力。最后利用提出的评估指标监控策略,动态调整模型权重的更迭,避免评估指标退化,提升模型训练效果。本文主要贡献如下。

(1)提出一个新的三维手部姿态估计网络,提取输入图片的二维特征,最大程度保留输入图片的手部特征信息,从数据质量角度增强后续 AK_GraFormer 模块的表现。

(2)使用具有残差连接的 res_AKGNN 图核自适应地学习不同节点的特征信息,增强模型的特征学习与表达能力,从而提高估计准确度。

(3)在训练过程中,设计基于评估指标监控的动态调整训练策略,根据训练情况冻结模型的敏感权重,在一定程度上能够避免过拟合问题,从而提升模型性能。

1 基于图卷积网络的三维手部姿态估计方法

基于图卷积网络,本文提出了一个新的手部姿态估计网络,本文模型整体框架如图 1 所示。首先,使用 Keypoint R-CNN 提取图片信息,得到图像视觉特征和二维位置信息。其次,将提取的信息输入到自适应核 AK_GraFormer 模块中,得到手部三维关键点坐标。AK_GraFormer 模块根据节点之间的连通性从图数据中提取局部特征和全局特征,自适应地控制反向传播梯度,挖掘输入数据在图结构下更复杂、更抽象的特征表示,增强网络的特征学习和表达能力。最后,使用设计的评估指标监控的动态训练策略,在训练过程中动态冻结 Keypoint R-CNN 模块的参数,稳定多模块协同优化过程,增强网络的整体训练效果。

1.1 特征信息提取

手部关节存在自相似性以及严重的遮挡情况,单一的二维位置只能捕捉平面信息,这导致网络在估计三维坐标时易产生歧义,所估计的三维手部姿态难以体现真实手部结构特性。鉴于此,本文在经典方法的基础上,将图像视觉特征和二维位置信息

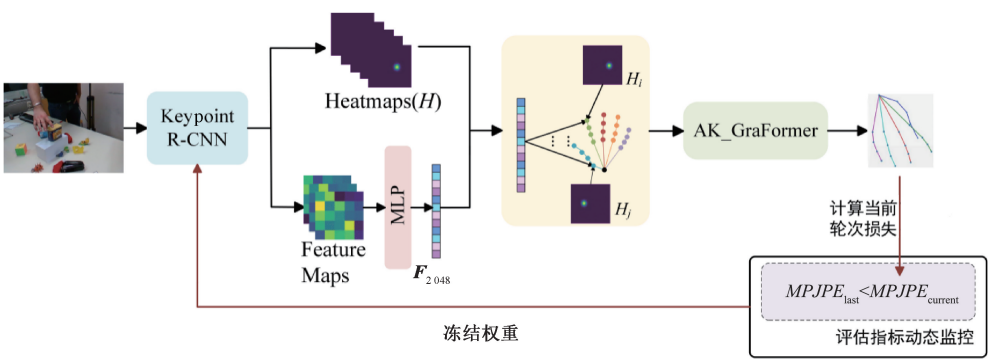

图 1 本文模型整体框架

Figure 1 Overall framework of model

同时作为后续模块的输入,旨在为网络保留更多三维空间的图像特征信息。

本文使用 Keypoint R-CNN 提取输入图片的特征信息,如图 1 所示。通过训练 Keypoint R-CNN 得到关键点的位置热图(Heatmaps(H))和多尺度特征图(Feature Maps),再借助特征提取器(MLP)对多尺度特征图实施非线性变换,生成压缩的 2 048 维特征向量($\boldsymbol{F}_{2\,048}$)。随后,这些特征向量被附加至热图上,形成一种新的表示形式,即图像视觉特征信息。在热图中,每个通道对应一个关键点的二维位置信息,而附加的特征向量则为各关键点提供了更为丰富的上下文信息。特征信息提取模块为 AK_GraFormer 提供了高质量输入数据,使得网络能够构建更为精准的手部节点表示。

1.2　AK_GraFormer 模块设计

1.2.1　AK_GraFormer 模块

在图神经网络中,节点在不同任务中对信息的需求层次不同。在手部节点构成的图结构中,靠近手掌的节点在整个图结构中起着核心的连接作用,需要模型提取更丰富、精细的信息以进行三维关键点估计;而靠近指尖的节点的位置与姿态更多是从与之相连的节点的关系来间接确定,对信息的需求层次较低。因此,本文提出 AK_GraFormer 模块,在 GraFormer 模块中引入了改进的 AKGNN 层,旨在解决模型中不同节点的信息需求层次不同的问题,增强模型的特征表达与学习,模块结构如图 2 所示。

图 2 中右半部分展示了 AK_GraFormer 模块的结构,该模块通过学习、表达特征信息,实现对手部节点的三维坐标估计。左半部分是 AK_GraFormer 模块中 Res_AKGNN 层的结构,通过决策手部关节

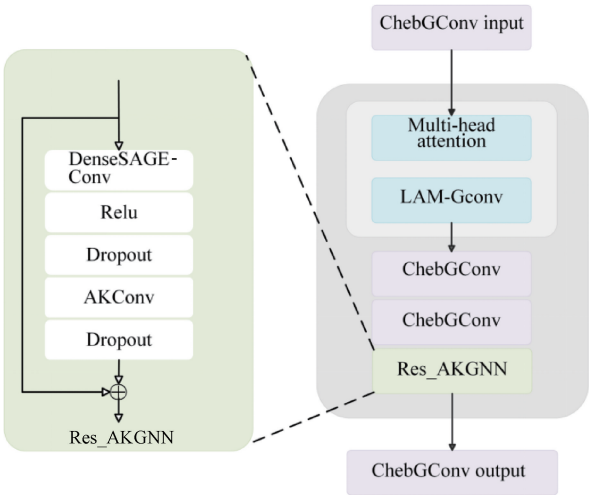

图 2　AK_GraFormer 模块图

Figure 2　Module diagram of AK_GraFormer

的不同信息层次需求以自适应保留更重要的特征数据。相比于 GraFormer 模块,改进后的 AK_GraFormer 模块能够更有效地处理手部节点图结构中的信息差异。它根据不同节点的位置和在图中的重要性,自适应地调整信息的传递和处理方式,避免了不必要的信息冗余和计算资源浪费,同时使得最终的预测结果保留更多有效的手部结构特征。

1.2.2　Res_AKGNN 层

Res_AKGNN 层的结构如图 2 左半部分所示,根据手部节点的特点,本文提出在 Res_AKGNN 层中加入 DenseSAGEConv,以根据节点的邻居信息更新节点的特征表示;利用 AKConv 层通过自适应核学习来调整图卷积的滤波器权重,以适应不同频率的信息;使用残差连接将经过传播和变换后的特征与原始输入相加,得到传播后的特征。而后通过全局读出函数对节点表示进行整合,利用多层感知机进行特征变换,得到保留更多信息的特征。

DenseSAGEConv 层通过有效地融合跨层级特征,形成一条密集的特征传递路径,进而以更有效的信息流提高模型的表征学习能力。对于每个节点,该层首先聚合其邻居节点的特征,随后通过权重矩阵和激活函数进行特征变换,最终将邻居信息整合到节点自身的嵌入向量中。

$$\boldsymbol{h}_i^{\text{new}} = \sigma\Big(\sum_{j\in N(i)} w_j \boldsymbol{h}_j + \boldsymbol{b}\Big)。\tag{1}$$

式中:$\boldsymbol{h}_i^{\text{new}}$ 为更新后的节点 i 的特征;$N(i)$ 为节点的邻居节点集合;w_j 为邻居节点 j 的权重;$\boldsymbol{h}_j$ 为邻居节点的特征;$\boldsymbol{b}$ 为偏置项;$\sigma(\cdot)$ 为激活函数。

进入 Res_AKGNN 层后,节点特征在传播过程中的信息提取过程表示如下:

$$\boldsymbol{Y}^P = \text{softmax}(f_{\text{MLP}}(\text{READOUT}(\boldsymbol{H}^{(k)})))。\tag{2}$$

式中:$\boldsymbol{H}^{(k)}$ 为节点特征在传播过程中生成多个中间节点 k 的表示矩阵;READOUT(·)为全局读出函数;f_{MLP} 为多层感知机。

输入特征经过 Res_AKGNN 层传播和变换后得到特征 $\boldsymbol{h}$,再通过残差连接将 $\boldsymbol{h}$ 与原始输入相加。残差连接使得梯度能够直接从深层网络传播回浅层网络,避免了在多层网络传播过程中梯度逐渐变小甚至消失的问题。

Res_AKGNN 层使得模型能够根据节点的实际需求选择合适的信息层次进行决策。在手部姿态估计中,对于一些关键关节即靠近根节点的关节,需要更多的全局信息来确定其准确位置,而对于一些末梢关节即如指间关节,局部信息就足够。AKGNN 能够全局读出函数,可以帮助模型更好地处理这种

差异。因此,Res_AKGNN 层对经过前面层处理后的特征数据进行特定的特征融合、变换或者信息传递等操作,增强了整个模型对基于图结构的输入数据的特征学习和表达能力,提高模型对于三维姿态估计的准确性。

1.3 评估指标动态监控模块

在模型训练阶段,通常会观察到训练过程呈现一定的收敛趋势,且相关评估指标处于持续向好的状态。然而,实际训练过程中可能会出现某一训练轮次(*Epoch*)下,评估指标突然变差的情况。在本模型实际训练过程中,每一轮训练结束后即刻验证模型效果。随着训练 *Epoch* 的增加,可以观察到验证结果中二维损失的增速明显大于三维损失。这是因为在训练过程中,KeyPoint R-CNN 的特征空间和 AK_GraFormer 模块的决策边界需要协同收敛。若 KeyPoint R-CNN 特征持续变化,AK_GraFormer 模块需不断适应新特征分布,增加了学习难度。此外,Vasconcelos 等[23]研究指出关键点检测器与后续模块联合训练时,阶段性冻结能提升最终精度。

针对这一现象,本文设计了一种评估指标监控动态训练策略,对 KeyPoint R-CNN 模块的模型权重采取动态冻结处理。达到一定条件后,该模块在后续的训练过程中保持固定,不再参与更新迭代。若当前 *Epoch* 验证集 $MPJPE_{\text{current}}$(当前轮次的平均每关节位置误差)大于上一个 *Epoch* 验证集 $MPJPE_{\text{last}}$ 时,则将上一个 *Epoch* 模型参数 $Model\ Parameters_{\text{last}}$ 赋值给当前 *Epoch* 模型参数 $Model\ Parameters_{\text{freezed}}$,并且将 KeyPoint R-CNN 模块的参数($Param \in Model\ Parameters_{\text{freezed}}^{\text{backboneUrpn}}$)冻结。因此,在后续训练中将不再计算该模块的梯度(*Param. requires_grad*)。本文通过这种方式避免异常训练轮次所带来的不良影响,保障模型训练的稳定性与有效性,进而提升手部姿态估计任务的整体性能表现。具体算法如下。

算法 1:评估指标动态调节算法。

输入:当前 *Epoch* 验证集 $MPJPE_{\text{current}}$、上一个 *Epoch* 验证集 $MPJPE_{\text{last}}$、上一个 *Epoch* 模型参数 $Model\ Parameters_{\text{last}}$;

输出:冻结权重后的参数 $Model\ Parameters_{\text{freezed}}$。

① if $MPJPE_{\text{current}} > MPJPE_{\text{last}}$ do
② $Model\ Parameters_{\text{freezed}} \leftarrow Model\ Parameters_{\text{last}}$;
③ for *Param* in $Model\ Parameters_{\text{freezed}}$ do
④ if $Param \in Model\ Parameters_{\text{freezed}}^{\text{backboneUrpn}}$ do
⑤ *Param. requires_grad* = False;
⑥ return $Model\ Parameters_{\text{freezed}}$。

1.4 损失函数

为了有效监督网络训练,缩小预测坐标值与真实坐标值的差距,本文使用两个损失函数。总损失定义为

$$L = \lambda_1 L_{2D} + \lambda_2 L_{3D}。 \tag{3}$$

式中:通过实验观察不同权重组合对验证效果的影响逐步调整权重值,最终权重值设置为 $\lambda_1 = 0.1$,$\lambda_2 = 0.9$;L_{2D} 为二维姿态估计损失;L_{3D} 为三维姿态估计损失。

$$L_{2D} = \frac{1}{n}\sum_{i=1}^{n} \| J_i - \hat{J}_i \|^2。 \tag{4}$$

式中:J_i 表示标注的真实的第 i 个关节的二维坐标;$\hat{J}_i$ 表示预测的第 i 个关节二维坐标。

$$L_{3D} = \frac{1}{n}\sum_{i=1}^{n} \| P_i - \hat{P}_i \|^2。 \tag{5}$$

式中:P_i 表示标注的真实的第 i 个关节的三维坐标;$\hat{P}_i$ 表示预测的第 i 个关节三维坐标。

2 实验结果及分析

2.1 数据集和评价方法

本文基于 HO-3D_v3 数据集[24]和 FreiHAND 数据集[25]进行实验。HO-3D_v3 数据集是一个手和物体在严重相互遮挡情况下的 3D 姿态注释的数据集。该数据集中的 68 个序列包含 10 个不同的人操作 10 个不同的物体,这些物体来自 YCB 物体数据集,包含 77 558 张图像的注释,这些图像被分为 66 034 张训练图像(来自 55 个序列)和 11 524 张评估图像(来自 13 个序列)。FreiHand 数据集于 2019 年发布,包含 32 个对象的手势。该数据集具有背景多样、视点多变、动作复杂等特点,其中部分手势是手部与物体交互的动作,包含 13 024 个训练样本和 3 960 个评估样本。

本文使用以下 3 个评估指标:平均每关节位置误差(*MPJPE*),通过计算预测和实际三维关节位置之间的欧几里得距离度量结果好坏;*PA-MPJPE* 为 *MPJPE* 的改良版,采用普氏分析对预测姿态与真实姿态进行刚性对齐(旋转、平移),消除全局位置与方向变化的影响,从而专注于评估关节间的相对姿态误差;检测关节点百分比(*PCK*)曲线下面积(*AUC*),其中 *PCK* 指正确 3D 关节的百分比。

2.2 实验环境与参数

本文实验环境:操作系统为 Ubuntu 2022.04 LTS、GPU 配置为 RTX-4090D、CPU 配置为 AMD EPYC 9754、内存大小为 60 GB。神经网络的实现框

架为 PyTorch 和 PyTorch Geometric 框架。

本文实验的主要参数:输入的手部按 Mesh2D 计算训练验证检测框数据,*train_batch* 大小设置为 32,*test_batch* 大小设置为 1,使用 Adam 作为优化器,初始学习率设置为 0.000 1。

2.3　可视化分析

本文模型在 HO-3D_v3 数据集上的可视化结果如图 3 所示。原图中的矩形框用来表示模型检测到的手部位置,而真实值和预测值中的手部姿态通过不同的线条颜色区分不同手指。通过可视化展示,能够清晰地看到模型预测结果与真实数据之间的对比。观察这些图像时,可以发现手部关节在未被遮挡的情况下模型能够准确地预测关节位置,与真实值几乎重合。然而在发生手部和物体遮挡的情况下,预测值与真实值存在一定的偏差,但预测的关节图符合天然的手部结构的生理约束。因人工标注数据方法的成本高,HO-3D_v3 数据集采用机器标注方法,标注存在一定误差。因此在训练过程中,标注数据甚至对模型的学习产生了负面影响。

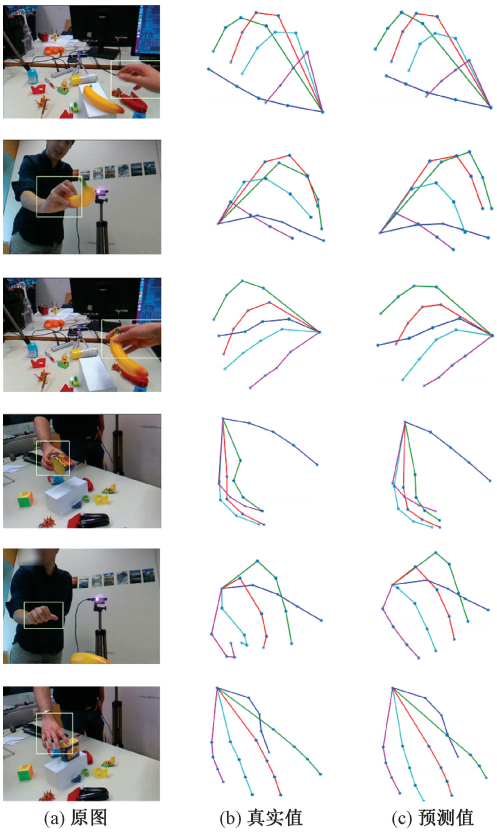

图 3　可视化结果

Figure 3　Visualization results

2.4　消融实验

本文在 HO-3D_v3 数据集上进行了消融实验,以验证本文所提出模块的有效性。在本文模型中,采用分别叠加信息提取模块、AK_GraFormer 模块和评估指标动态监控模块进行实验。消融实验结果如表 1 所示。其中在加入评估指标动态监控模块之前,模型训练过程中出现的验证损失值变化情况如图 4 所示。从图 4 中可以清晰地看出,2D 验证损失随着训练轮次的增加而呈现上升趋势,模型对验证数据的拟合效果变差,其预测的准确性会随之降低,这对训练的影响是负面的。

表 1　各模块消融实验

Table 1　Ablation study for each module

名称	*MPJPE*/mm	*PA-MPJPE*/mm	*AUC*
基线模型[13]	25.60	11.20	0.755
基线模型+信息提取模块	24.29	9.88	0.798
基线模型+信息提取模块+AK_GraFormer 模块	23.64	10.02	0.793
本文模型	22.36	9.45	0.812

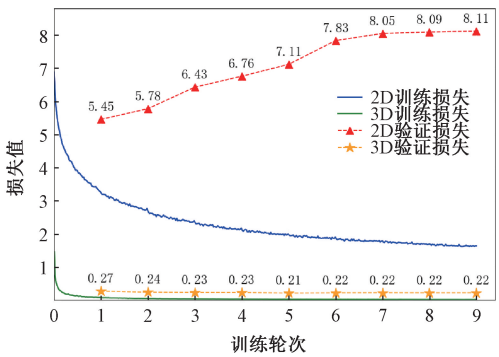

图 4　损失值变化情况

Figure 4　Changes in value of losses

从表 1 实验结果来看,基线模型误差较高,反映了其特征表达单一、关节间的拓扑结构未被充分挖掘。在加入融合信息提取模块后,模型对姿态的上下文感知能力变强,减少了因视角变化和遮挡等引起的相对姿态误差。而在此基础上加入 AK_GraFormer 模块后,*MPJPE* 有所下降,但 *PA-MPJPE* 却略微上升。改进的 res_AKGNN 层强化了网络的复杂姿态表达能力,从而降低了 *MPJPE*。但是两模块联合训练,导致特征漂移与参数更新异步性,最终影响模型精度。加入本文提出的全部模块后,模型误差显著降低,这表明评估指标动态监控模块的加入及时地调整了模块训练策略,确保各模块协同优化。

消融实验的可视化结果如图 5 所示,在手部姿态复杂和存在遮挡的情况下,本文方法依旧表现出良好的性能,估计出准确且符合手部结构特点的手部姿态。

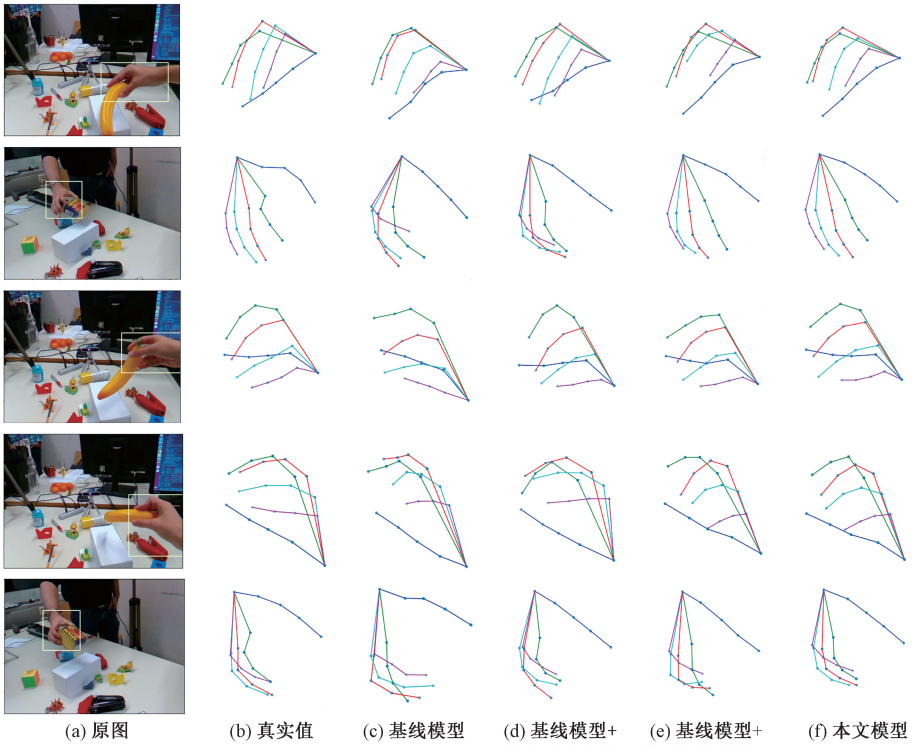

图 5 消融实验可视化结果

Figure 5 Visualization results of ablation study

2.5 定量比较

表 2 比较了本文提出的模型与其他模型在 HO-3D_v3 数据集上的性能。其中,为保证表格的整洁性,将杨冰等[26]提出的基于改进 FastMETRO 的三维手部姿态估计方法称为 FastMETRO。从评估的结果得知,本文模型在 *PA-MPJPE* 评估指标上取得最好结果。表 3 为本文提出的模型与其他模型在 FreiHand 数据集上的比较结果,本文模型在所比较指标上的结果均为最优。尽管同类方法在重度遮挡场景下的手部姿态估计中展现出较低的平均关节位置误差,但却存在着手部拓扑结构失真的问题。而本文方法能够通过学习图像特征信息优化图结构学习,提高姿态估计的准确度,同时使生成的手部姿态符合人手结构特点。

表 2 HO-3D_v3 数据集的定量比较

Table 2 Quantitative comparison of HO-3D_v3 datasets

模型名称	*PA-MPJPE*/mm	*AUC*
S^2HAND[27]	11.50	0.769
ArtiBoost[28]	10.80	0.785
PyMAF-X[29]	10.80	—
HMP[30]	10.10	—
FastMETRO[26]	10.50	—
本文模型	9.45	0.812

表 3 FreiHand 数据集的定量比较

Table 3 Quantitative comparison of FreiHand datasets

模型名称	*PA-MPJPE*/mm	*AUC*
I2UV-HandNet[31]	6.70	0.856
Mesh graphormer[32]	5.90	0.874
HandMIM[33]	6.29	0.871
HaMeR[34]	6.00	—
本文模块	5.78	0.878

3 结论

本文针对手部姿态估计提出基于图卷积的三维手部姿态估计算法。为了减少中间的二维位置对最终的三维位置的影响,同时为姿态估计网络提供更多辅助信息,本文通过提取输入图片的特征信息,得到图像视觉特征和二维位置信息;将提取的信息输入到 AK_GraFormer 模块中,引入改进的 res_AKGNN 层,自适应学习输入数据更深层的特征表示,提高模块的表达能力,使得估计结果更精准;设计评估指标动态监控模块,以应对训练过程中评估指标退化现象,避免过拟合,有效提升了模型性能。在 HO-3D_v3 数据集与 FreiHand 数据集上进行实验,结果表明:在单张彩色图片手部三维姿态估计任务中,所提方法相比其他同类方法具有明显优势,刚性对齐后的平均每关节位置误差(*PA-MPJPE*)最高

降低了 12.50 百分点,检测关节点百分比曲线下面积(*AUC*)最高提高了 3.44 百分点。

参考文献:

[1] Sridhar S, Feit A M, Theobalt C, et al. Investigating the dexterity of multi-finger input for mid-air text entry[C]// Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. New York: ACM, 2015: 3643-3652.

[2] Oikonomidis I, Kyriazis N, Argyros A A. Tracking the articulated motion of two strongly interacting hands[C]// Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2012: 1862-1869.

[3] Tkach A, Pauly M, Tagliasacchi A. Sphere-meshes for real-time hand modeling and tracking[J]. ACM Transactions on Graphics, 2016, 35(6): 1-11.

[4] Romero J, Tzionas D, Black M J. Embodied hands: modeling and capturing hands and bodies together[J]. ACM Transactions on Graphics, 2017, 36(6): 1-17.

[5] Pavlakos G, Choutas V, Ghorbani N, et al. Expressive body capture: 3D hands, face, and body from a single image[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 10967-10977.

[6] Keskin C, Kirac F, Kara Y E, et al. Hand pose estimation and hand shape classification using multi-layered randomized decision forests[C]// Proceedings of the 12th European conference on Computer Vision. New York: ACM, 2012: 852-863.

[7] 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.

[8] Pan Xiaoying, Li Shoukun, Wang Hao, et al. LGCANet: lightweight hand pose estimation network based on HRNet[J]. The Journal of Supercomputing, 2024, 80(13): 19351-19373.

[9] Hoang D C, Xuan Tan P, Pham D L, et al. Efficient multimodal fusion for hand pose estimation with hourglass network[J]. IEEE Access, 2024, 12: 113810-113825.

[10] Zhan Zhi, Luo Guang. Multiscale feature fusion network for monocular complex hand pose estimation[J]. Electronics Letters, 2023, 59(24): e13044.

[11] Panteleris P, Oikonomidis I, Argyros A. Using a single RGB frame for real time 3D hand pose estimation in the wild[C]//Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE, 2018: 436-445.

[12] Doosti B, Naha S, Mirbagheri M, et al. HOPE-Net: a graph-based model for hand-object pose estimation[C]// Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 6607-6616.

[13] Zhao Weixi, Wang Weiqiang, Tian Yunjie. GraFormer: graph-oriented transformer for 3D pose estimation[C]// Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 20406-20415.

[14] Li Zhixin, Shang Fanqi, Huan Zhan, et al. Human activity recognition based on hybrid feature graph convolutional neural network[J]. Journal of Zhengzhou University (Engineering Science), 2024, 45(4): 46-52. [李志新, 商樊淇, 郇战, 等. 基于混合特征图卷积神经网络的人体行为识别方法[J]. 郑州大学学报(工学版), 2024, 45(4): 46-52.]

[15] Cai Yujun, Ge Liuhao, Liu Jun, et al. Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks[C]// Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 2272-2281.

[16] Aboukhadra A T, Malik J, Robertini N, et al. ShapeGraFormer: GraFormer-based network for hand-object reconstruction from a single depth map[J]. IEEE Access, 2024, 12: 124021-124031.

[17] Zhuang Nan, Mu Yadong. Joint hand-object pose estimation with differentiably-learned physical contact point analysis[C]//Proceedings of the 2021 International Conference on Multimedia Retrieval. New York: ACM, 2021: 420-428.

[18] Zhang Maomao, Li Ao, Liu Honglei, et al. Coarse-to-fine hand-object pose estimation with interaction-aware graph convolutional network[J]. Sensors, 2021, 21(23): 8092.

[19] Ma Shenglei, Li Jinghua, Kong Dehui, et al. 3D hand pose estimation based on double branches with multi-scale attention[J]. Chinese Journal of Computers, 2023, 46(7): 1383-1395. [马胜蕾, 李敬华, 孔德慧, 等. 基于双分支多尺度注意力的手三维姿态估计[J]. 计算机学报, 2023, 46(7): 1383-1395.]

[20] Yang Wenji, Xie Liping, Qian Wenbin, et al. Coarse-to-fine cascaded 3D hand reconstruction based on SSGC and MHSA[J]. The Visual Computer, 2025, 41(1): 11-24.

[21] He Kaiming, Gkioxari G, Dollár P, et al. Mask R-CNN[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 2980-2988.

[22] Ju Mingxuan, Hou Shifu, Fan Yujie, et al. Adaptive kernel graph neural network[C]//Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22). Washington D C: AAAI, 2022: 7051-7058.

[23] Vasconcelos C, Birodkar V, Dumoulin V. Proper reuse of image classification features improves object detection[C]//Proceedings of the 2022 IEEE/CVF Conference on

Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 13618-13627.
[24] Hampali S, Sarkar S D, Lepetit V. HO-3D_v3: improving the accuracy of hand-object annotations of the HO-3D dataset[PP/OL]. V1. arXiv (2021-07-02)[2025-12-19]. https://doi.org/10.48550/arXiv.2107.00887.
[25] Zimmermann C, Ceylan D, Yang Jimei, et al. FreiHAND: a dataset for markerless capture of hand pose and shape from single RGB images[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 813-822.
[26] Yang Bing, Xu Chuyang, Yao Jinliang, et al. 3D hand pose estimation method based on monocular RGB images [J]. Journal of Zhejiang University (Engineering Science), 2025, 59(1): 18-26. [杨冰, 徐楚阳, 姚金良, 等. 基于单目RGB图像的三维手部姿态估计方法[J]. 浙江大学学报(工学版), 2025, 59(1): 18-26.]
[27] Chen Yujin, Tu Zhigang, Kang Di, et al. Model-based 3D hand reconstruction via self-supervised learning[C]// Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 10446-10455.
[28] Yang Lixin, Li Kailin, Zhan Xinyu, et al. ArtiBoost: boosting articulated 3D hand-object pose estimation via online exploration and synthesis[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 2740-2750.
[29] Zhang Hongwen, Tian Yating, Zhang Yuxiang, et al. PyMAF-X: towards well-aligned full-body model regression from monocular images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(10): 12287-12303.
[30] Duran E, Kocabas M, Choutas V, et al. HMP: hand motion priors for pose and shape estimation from video [C]//Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE, 2024: 6341-6351.
[31] Chen Ping, Chen Yujin, Yang Dong, et al. I2UV-HandNet: image-to-UV prediction network for accurate and high-fidelity 3D hand mesh modeling[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 12909-12918.
[32] Lin K, Wang Lijuan, Liu Zicheng. Mesh graphormer [C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 12919-12928.
[33] Liu Zuyan, Lin Gaojie, Wang Congyi, et al. HandMIM: pose-aware self-supervised learning for 3D hand mesh estimation[PP/OL]. V1. arXiv (2023-07-29)[2025-12-19]. https://doi.org/10.48550/arXiv.2307.16061.
[34] Pavlakos G, Shan Dandan, Radosavovic I, et al. Reconstructing hands in 3D with transformers[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2024: 9826-9836.

3D Hand Pose Estimation Based on Graph Convolution Network

PENG Chunyan[1,2], WANG Xuan[1,2], CHEN Yangbo[1,2], HE Gangbo[1,2]

(1. College of Computer, Qinghai Normal University, Xining 810016, China; 2. The State Key Laboratory of Tibetan Intelligence, Qinghai Normal University, Xining 810016, China)

Abstract: In the task of 3D hand pose estimation from a single image in color, challenges such as occlusion and high self-similarity of hand parts might lead to large prediction errors and unnatural hand structures. To address these issues, a graph convolution-based 3D hand pose estimation method was firstly proposed. Visual features and 2D keypoint positions were extracted from the input image using Keypoint R-CNN. These features were then fed into an improved adaptive kernel graph convolution module (AK_GraFormer). Subsequently, a residual-connected AKGNN graph kernel was introduced to adaptively process graph-structured data, thereby enhancing the model's feature learning and representation. Finally, a dynamic training strategy was employed, which was monitored by a proposed evaluation metric, to optimize estimation performance. Experimental results on the HO-3D_v3 and FreiHand datasets demonstrated that the proposed method outperformsed existing approaches in monocular 3D hand pose estimation. Specifically, the procrustes-aligned mean per joint position error (*PA-MPJPE*) was reduced by up to 12.50 percentage points, and the area under the curve (*AUC*) of the percentage of correct keypoints metric was improved by up to 3.44 percentage points compared to state-of-the-art methods.

Keywords: 3D hand pose estimation; graph convolution networks; feature extraction; optimisation of graph kernel learning; dynamic adjustment of assessment indicators