基于双分支多尺度注意力的手三维姿态估计

马胜蕾 李敬华 孔德慧 王立春 王少帆 尹宝才
(北京工业大学信息学部多媒体与智能软件技术北京市重点实验室 北京 100124)

摘 要 手部三维姿态估计旨在基于输入的二维手势图像预测手的三维关节点位置,其在虚拟现实、自然人机交互、自动驾驶等领域有广泛的应用前景.基于单张彩色图像的手姿态估计具有现实应用的普适性但也面临更大挑战.针对手部多关节复杂难建模问题,本文提出了基于双分支的手三维姿态估计框架.所提双分支网络结构的一个分支用于描述同一手指不同关节之间的物理连接关系,另一分支用于描述不同手指相同关节之间的对称运动关系,两种结构互补建模了手关节之间的复杂关联关系.对于每一分支,提出了基于多尺度注意力GUnet和改进语义图卷积的单分支手姿态估计方法,利用手关节的多尺度上下文信息和尺度间注意力提升姿态估计的准确性.本文在公开的STB和FreiHand数据集上进行了系列实验,实验结果表明本文所提方法优于现有的基于单张RGB图像的手姿态估计方法,平均关节误差相对基线方法分别降低了0.6 mm和0.8 mm.

关键词 手姿态估计;双分支;图卷积;多尺度;注意力机制
中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2023.01383

3D Hand Pose Estimation Based on Double Branches with Multi-Scale Attention

MA Sheng-Lei LI Jing-Hua KONG De-Hui WANG Li-Chun WANG Shao-Fan YIN Bao-Cai
(*Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Faculty of Information Technology, Beijing University of Technology, Beijing* 100124)

Abstract 3D hand pose estimation aims to regress 3D location of the hand joints based on the input 2D image, which has been widely applied to virtual reality, natural human-computer interaction, automatic driving and other fields. The hand pose estimation from a single RGB image is more pervasive but faces many challenges. To model complex hand joints, this paper proposes a kind of 3D hand pose estimation method based on double branches. In the proposed double branches network structure, one is used to describe the classical physical connection among the joints of the same finger, and the other is used to describe symmetrical connection among the corresponding joints of different fingers. The complementary topology effectively models complex associations among the joints. With regard to each branch, this paper presents a kind of multi-scale attention for hand pose regression which improves the hand pose estimation accuracy via multi-scale representation and scales attention. Experimental results on STB and FreiHand datasets show that the proposed method is superior to existing hand pose estimation methods, and the average joint error is respectively reduced by 0.6 mm and 0.8 mm relative to the baseline.

Keywords hand pose estimation; double-branches; graph convolution; multi-scale; attention mechanism

收稿日期:2022-02-09;在线发布日期:2023-01-06.本课题得到国家自然科学基金(62172022,U21B2038,61876012)、北京市自然科学基金(4202003)资助.**马胜蕾**,硕士,主要研究方向为图像处理、机器学习.E-mail:msl_qing@163.com.**李敬华**,博士,副教授,硕士生导师,中国计算机学会(CCF)会员,主要研究领域为图像处理、机器学习.**孔德慧**,博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为计算机图形学、计算机视觉、虚拟现实等.**王立春**,博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为人工智能、人机交互.**王少帆**,博士,副教授,硕士生导师,中国计算机学会(CCF)会员,主要研究领域为模式识别、机器学习.**尹宝才**(通信作者),博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为多媒体技术、跨媒体智能、视频编码.E-mail:ybc@bjut.edu.cn.

1 引 言

手部三维姿态估计的任务是从输入的手部图像预测手关节点的三维位置,其在虚拟现实、增强现实、人机交互、自动驾驶等领域有广阔的应用前景.不过手姿态估计也面临诸多挑战性问题.第一,手是多关节连接体,自由度多,形状变化复杂;第二,手指的运动变化既独立又相互关联,可以表达丰富的语义信息,难以理解;第三,双手手部姿态往往具有联动和协同,进一步增加了手势姿态估计的难度;第四,目前的二维图像获取手段难以获得手部姿态的精准三维位置信息.综上所述,手部三维姿态估计研究具有重要的理论研究意义和实际应用价值.

手部三维姿态估计的分类方法很多,可分为基于传统的机器学习和计算机视觉技术的方法以及基于深度学习的方法;传统方法通常提取手工定义特征并通过随机森林等方法估计手部姿态.基于深度学习的方法可分为基于检测的方法和基于回归的方法.基于检测的方法首先生成每个关节的概率密度图即热图,然后计算热图上的最大值获得每个关节的确切位置.基于回归的方法旨在通过深度网络直接预测各关节点的位置.基于深度学习的方法也可根据输入图像类型分类,分为基于深度图像、基于多张彩色图像、基于单张彩色图像的方法、基于 RGBD 图像的方法以及基于多模态训练单模态测试的方法[1],其中,前三种是主流方法.

基于深度图像进行手部三维姿态估计的方法受限于深度相机的采集环境要求和有限的成像范围,深度图像质量往往不高;并且虽然深度传感器的价格已达到消费级水平,但相对彩色图像,深度图像的获取还不普及.目前,基于深度图像估计手三维姿态的应用主要还局限于一些娱乐性游戏以及科学实验场景等.基于多张彩色图像估计手的三维姿态主要是基于立体视觉原理进行三维重建.不过同步采集多张不同视角的图像需要布控多视角图像同步采集环境,这在现实应用中必然受限.由于单张彩色图像获取的普适性,基于单张彩色图像的手三维姿态估计最具实际应用潜力,不过由于彩色图像中手指各关节外观的相似性,以及部分手势动作所表现出的手关节间遮挡以及广泛的手物交互应用中的物体遮挡等因素,基于单张彩色图像估计手的三维姿态是非常具有挑战性的问题.

鉴于单张彩色图像获取的便捷性,本文拟研究基于单张彩色图像的手部三维姿态估计.现有的基于单张彩色图像预测手部三维姿态的方法可分为单阶段法和两阶段法.单阶段法直接从图像回归手部各关节的三维位置,没有充分利用手关节之间的空间关系,预测精度不高.两阶段法最初被用于人体姿态估计[2-3],Zimmermann 等人[4]在 2017 年首次将两阶段法引入到基于单张彩色图像的手姿态估计,该方法是目前普遍应用的方法.第一个阶段检测手的二维关节点位置,即对输入图像,首先进行特征提取,检测手的二维关节点位置.第二阶段从手的二维关节点位置回归手的三维关节点位置.受上述工作启发,本文采用两阶段法开展该研究工作.因此精准的二维姿态估计以及二维姿态到三维姿态的预测都是重要的.其中,二维关节点的精准检测是三维姿态估计的基础.手部各关节之间是有关联关系的,利用好手关节间的强关联性实现精准的二维姿态估计以利于后续的三维姿态估计是本文的第一个解决思路.二维姿态到三维姿态的预测是高度非线性问题,本身是严重的病态问题,是有难度的.现有方法或者通过优化方法或者通过卷积神经网络、图卷积神经网络等学习方法进行预测,如何从二维关节点挖掘更多信息以准确预测三维姿态是本文工作的第二个着力点.

由于手指各关节之间存在连接关系,用好手的结构信息是解决上述问题的一个可行思路.在手部姿态估计任务中引入手部的结构信息必然有助于准确建模手关节之间的复杂关联关系.最容易想到的是手指各关节之间的物理连接关系,Guo 等人[5]引入手的物理结构信息,从手的二维姿态估计三维姿态.不过传统的物理结构连接建模了手指各关节之间的关联性,但没有利用不同手指的相应关节之间的关系.为此本文提出基于物理和对称结构约束的双分支手姿态估计框架,可描述手指内和手指间的关联.

对于双分支结构中的每一个分支,参考近年来手部三维姿态估计的最新研究成果[6-7],本文基于图卷积神经网络进行二维和三维关节点预测.特别地,鉴于图 U-net[8]在分类、人体姿态估计[9]等领域取得的巨大成功,本文引入图 U-net 作为基线网络进行手部三维姿态回归.传统图 U-net 网络仅基于解码器的最后一层特征作为表示层进行回归,没有利用各关节的多尺度特征.为了整合进各关节的不同感受野的特征,本文工作融合了解码器不同层的特征图预测手部各关节点的三维位置.同时,考虑到各尺度特征图对最终预测的贡献不同,本文引入注意力机制学习每个尺度的特征图对手部三维姿态估计的贡献.在每一分支,基于多尺度注意力机制的手姿态估计旨在充分挖掘各关节的更有表现力特征,从

而提高手姿态估计的准确性.

本文的主要贡献包括如下两个方面：

(1) 提出了基于物理和对称结构约束的双分支手姿态估计框架. 物理和对称双分支结构互补建模了手关节的复杂关联关系，从而提升了手三维姿态估计的准确性.

(2) 提出了基于多尺度注意力 GUnet 和改进语义图卷积的单分支手姿态估计方法，设计尺度间注意力融合基于改进 GUnet 提取的手关节在不同尺度的上下文特征，提升了基于手的二维姿态回归三维姿态的性能.

2　相关工作

现有的手部三维姿态估计方法可分为基于模型的方法、基于学习的方法和混合的方法. 近年来，随着机器学习和深度学习的快速发展，基于学习的方法成为研究的主流. 根据输入图像的类型不同，可以将手姿态估计分为三类：基于深度图像的手姿态估计、基于多张 RGB 图像的手姿态估计和基于单张 RGB 图像的手姿态估计. 接下来本节将分别阐述每类方法.

基于深度图像的手姿态估计：深度图像因包含深度信息，通常被称作 2.5 维图像，因此基于深度图像估计手的三维姿态是目前的主流方法. Tompson 等人[10]首次将卷积神经网络应用到三维手部姿态估计，不过卷积神经网络只是用于特征提取，而三维姿态估计通过逆向运动学实现. Wan 等人[11]提出一种稠密的像素级姿态估计方法，使用堆叠沙漏网络[12]从给定的深度图像估计二维热图、三维热图和三维单位矢量场. 三维热图与三维单位矢量场相结合，有效地构成了二维深度图像像素与手部关节之间的三维偏移向量，并由此推断三维手部关节点的坐标. Du 等人[13]提出一种交叉信息网，该网络将手部姿态估计任务分解为手掌姿态估计子任务和手指姿态估计子任务，采用两分支交叉连接结构在子任务之间共享有益的补充信息，利用两个分支分别回归手掌姿态和手指姿态. Moon 等人[14]为了充分利用深度图像的三维信息，把深度图转换为体素表示，提出基于体素到体素预测的网络结构，估计每个体素是手关节点的可能性. 虽然目前基于深度图像的手部姿态估计已经取得了比较好的进展，但是由于现阶段的深度摄像机的成像范围十分有限，而且质量不够高，限制了依赖深度图像作为输入进行手部姿态估计的发展；同时由于深度图像相对彩色图像更难获取，所以基于深度图像的手姿态估计在实际中的应用不多.

基于多张 RGB 图像的手姿态估计：相对于深度图像，彩色图像更易获取. 由于不同视角的 RGB 图像或者视频序列包含丰富的三维信息，所以基于多张 RGB 图像估计手的三维姿态是自然的想法. Cai 等人[7]基于图建模手的结构信息，进而提出基于图卷积的三维姿态估计方法挖掘二维手关节点序列蕴含的空间相关性和时间一致性. 他们提出一个局部到全局的网络架构，可有效捕获不同尺度上的图特征. 同时，该方法引入了一种非均匀图卷积策略，即对不同的邻接节点学习不同的卷积核权重. 基于多张 RGB 图像的手部姿态估计因不同视角数据的信息互补而能有效解决手部遮挡的问题，从而能够得到较高的预测精度，但是此方法所需要的训练集、测试集资源较大，并且同步获取同一手势动作的多张图像是困难的.

基于单张 RGB 图像的手姿态估计：与上述两种手姿态估计方法相比，基于单张 RGB 图像的姿态估计更实用. 不过，研究基于单张 RGB 图像的三维手部姿态估计更具挑战性. 因为缺乏深度信息，必然存在深度模糊问题. Zimmermann 等人[4]最先提出基于学习的方法从单张 RGB 图像估计手部姿态，其设计了由三个不同模块组成的深度学习网络框架. 使用 HandSegNet 网络对图像进行分割；使用 PoseNet 网络进行二维姿态估计；使用 PosePrior 和 Viewpoint 分别估计三维姿态和视点. Baek 等人[15]提出利用参数化手模型拟合输入 RGB 图像的方法. 该方法通过三维监督、手分割蒙版和二维监督来训练手网格估计器和绘制器，并使用梯度下降法进一步细化初始三维网格估计. Ge 等人[16]提出一种使用单张 RGB 图像估计三维手形状和姿态的方法，该框架在有标注的合成数据集上以有监督的方法训练模型，在没有三维手姿态标注的真实数据集上对模型进行微调. 该方法使用图卷积神经网络来重建一个完整的手部三维网格，再利用三维网格回归手的姿态. 该方法也引入深度图弱监督基于手三维姿态渲染的深度图.

综上，基于单张彩色图像的手姿态估计有更广阔的应用前景，不过上述方法并没有充分挖掘手关节之间的复杂关联关系. 为了解决上述问题，本文提出基于双分支多尺度注意力的手姿态估计网络.

3　方　法

本节首先介绍基于双分支的手三维姿态估计网

络的整体框架;然后介绍基于多尺度注意力的单分支手姿态估计方法;最后介绍网络使用的损失函数.

3.1 基于双分支的手三维姿态估计框架

本文提出了基于双分支多尺度注意力的手三维姿态估计网络(Double Branches with Multi-Scale Attention,DBMSA-Net),整体框架如图 1 所示.根据图 1 可见,输入是单张彩色图像,输出是两个分支的三维姿态估计结果的融合值.两个分支分别基于手关节之间的不同连接关系进行预测.对于每一个分支,手三维姿态估计方法包括两个阶段:手二维姿态检测和手三维姿态回归.具体地,对于输入的单张 RGB 图像,首先通过 ResNet50 提取图像特征,接下来通过一个全连接层(Fully Connected,FC)预测手部关节的初始二维坐标;然后利用不同的图结构分别建模同一手指不同关节之间的物理连接关系和不同手指相同关节之间的对称连接关系,进而进行精确的二维姿态估计和三维姿态回归.根据图 1,图卷积神经网络(Graph Convolutional Networks,GCN)是二维姿态估计和三维姿态回归的核心技术.构建图结构是 GCN 的基础,常用方法包括预先定义的方法以及自学习图结构的方法.本文经过对手关节的固有物理连接属性和对称运动属性的分析,提出基于两种预定义的图结构互补描述手关节的不同连接关系,两个分支的二维姿态估计和三维姿态回归分别基于各自预定义的图结构进行图卷积操作,进而学习得到手关节之间的连接权重和每个关节的特征.为区分这两个分支,本文将用于精确二维姿态估计的两个分支网络分别称为基于物理连接关系的图卷积网络(Physical-GCN,P-GCN)和基于对称连接关系的图卷积网络(Symmetrical-GCN,S-GCN).为了实现二维姿态到三维姿态的准确回归,提出基于多尺度注意力的图 U-net(Graph for U-net,GUnet)网络(Multiscale Attention GUnet,MSA-GUnet).关于两个分支预测结果的融合,本文分别采用了计算两个分支预测结果平均值的融合策略以及基于学习权重的融合策略.

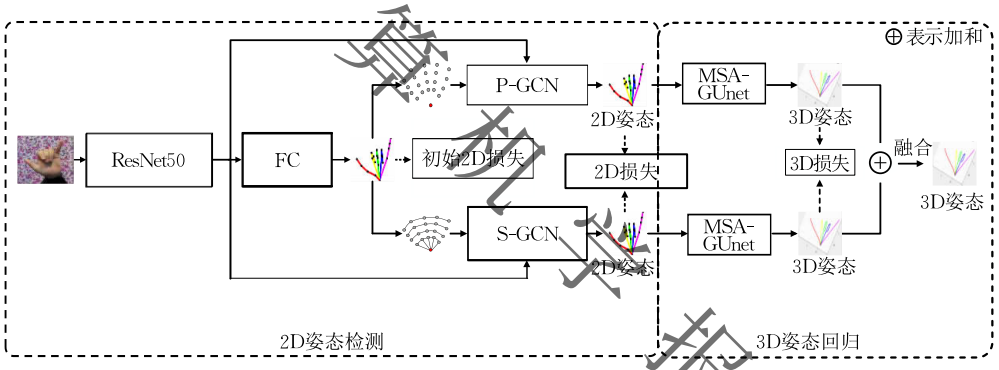


图 1 基于双分支的手姿态估计框架图

手部三维姿态估计旨在从输入图像估计一组预定义的手部关节点的三维位置.关于手部关节点模型的定义目前尚未达成共识,常见的手模型有如图 2 所示的三种定义:14 个关节的手模型、16 个关节以及 21 个关节的手模型,其中 21 个关节是使用最多的手模型,所以本文使用最为流行的 21 个关节的手模型[7,17].如图 2(c)所示,21 个关节的手模型包含一个腕关节以及每个手指的 4 个关节.通常地,食指、中指、无名指和小指的每个手指均有 4 个关节;而对于大拇指,显然少一个关节.为了计算上的方便,模型定义时增设了辅助关节,从而保证每个手指有 4 个关节.

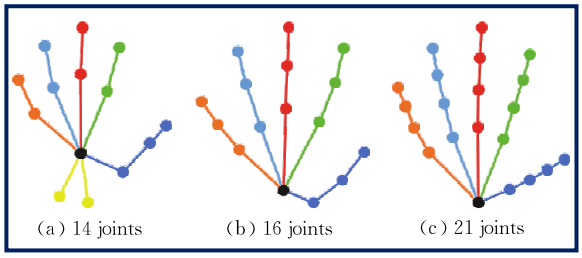
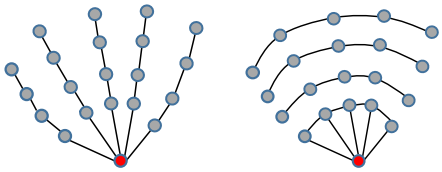


图 2 不同数量的手关节模型结构[18]

图 2 所示的关节间连接方式对应于手指各关节的物理连接关系,不同于同一手指不同关节之间显式存在的物理连接关系,对于部分手势而言,不同手指相同关节之间存在一致运动的性质,本文称其为对称关系,显然,这种关系是隐式存在的.以握拳手势为例,不同手指的相同关节具有一致的弯曲动作.为此,本文定义不同手指同一关节之间的对称连接关系以互补建模单纯物理连接关系不能精准表达的关节之间的关系.为此,本文设计了两种拓扑结构,如图 3 所示,第一种拓扑结构即传统的物理连接,第二种拓扑结构称为对称连接,这种对称连接描述了不同手指相同关节之间的联系.在网络结构中,这两种拓扑结构分别使用不同的邻接矩阵表示.



(a) 物理连接　　(b) 对称连接

图 3　手的两种拓扑结构

图 4 可视化了基于手关节之间的物理连接关系和对称连接关系预测两个手势的结果，不难看出，输入端的第一个手势更适合用对称分支预测，而第二个手势更适合用物理分支预测. 可见，同一手指不同关节之间以及不同手指同一关节之间都存在相互作用及空间联系，因此，本文提出物理和对称连接两种拓扑结构互补建模手关节之间的复杂关系.

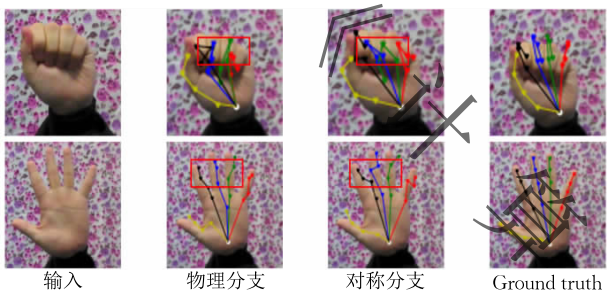


图 4　物理分支和对称分支的预测结果可视化

关于双分支的融合，本文参考文献[19]的做法，设计实现了两种融合策略. 一种是对两分支的预测结果求平均，另一种是基于学习的权重进行加权融合. 前者是对每一分支预测得到的 21 个关节的三维位置直接计算均值，后者的计算如式(1)：

$$\hat{y}_i=\omega_i^p\hat{y}_i^p+\omega_i^s\hat{y}_i^s+b_i \tag{1}$$

这里，$\hat{y}^p,\hat{y}^s\in\mathbb{R}^{3J}$ (J 为关节数)分别表示物理分支和对称分支的预测结果. ω_i^p,ω_i^s 为物理分支和对称分支估计姿态的第 i 关节的权重，b_i 是偏置. 参考文献[19]的做法，本文使用一个局部连接层来拟合权重，对于每个关节在三维空间位置的每一维度，独立学习权重.

3.2　基于多尺度注意力 GUnet 的单分支手姿态估计

前述手的双分支结构互补建模了手关节之间的复杂关联关系，有助于提升手三维姿态估计的准确性. 不过我们也看到，不同手势可能适合不同的连接关系建模，如图 4 的握拳动作，基于物理连接关系的姿态估计结果不理想，而在双分支融合预测框架中，每一分支都会影响预测结果，所以提升每一单分支的手姿态估计结果是必要的. 本文在每一单分支提出 MSA-GUnet 提取关节的多尺度上下文注意力特征从而更好回归手的三维位置. 如图 1 所示，手的三维姿态估计的每个分支包括两个部分：二维姿态检测和三维姿态回归；本节工作旨在完成第二部分任务即从二维姿态到三维姿态的回归.

本文所提出的 MSA-GUnet 如图 5 所示，根据图 5，所提方法采用了基线方法 Hope-net[20] 中的 GUnet[8] 框架，即改进的 Unet 框架[21]. 不过，在基线方法中，手三维姿态估计只利用 GUnet 解码子网的最后一层特征进行预测，没有利用各关节的多尺度特征. 为了整合各关节不同尺度的上下文特征，本文工作融合了解码器不同层的特征图即多尺度特征图预测手部各关节点的三维位置. 由于 GUnet 结构中解码器各层的输出维度不同，因此分别基于图卷积方法进行了上采样处理. 鉴于简单融合各尺度特征

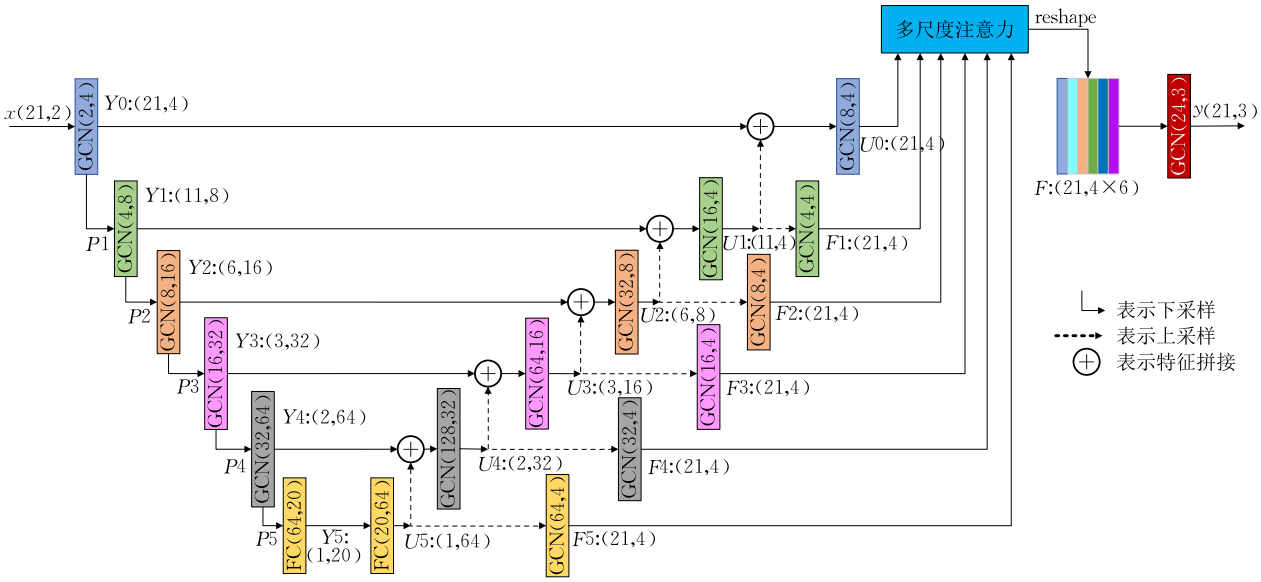


图 5　MSA-GUnet 网络结构

忽略了不同尺度特征对手姿态估计结果的重要性，因此本文引入注意力机制，提出基于多尺度注意力方法融合各尺度特征.接下来，本文将重点介绍基于GUnet的多尺度特征提取方法，以及基于多尺度注意力GUnet的手姿态估计方法.在GUnet中，非常核心的模块是图卷积，本文提出基于改进语义图卷积的方法.

多尺度表达即多分辨率表达，通常指用不同感受野的卷积核卷积图像所得的一系列特征图，多尺度特征图通常通过一个三阶张量表示.因为不同尺度编码不同尺寸的局部上下文，因此多尺度特征图张量具有很好的特征描述能力，已成功应用于各种计算机视觉问题[22-23].本文利用GUnet提取手关节的特征，输入是21个关节的二维特征，通过编码子网逐步得到手势的简化表示，即逐层近似减半的关节数量，以及每个关节逐步增大感受野的上下文特征；相对地，解码阶段，关节数近似两倍地增大，而特征数逐层减半.这里，通过拼接编码器对称层，融合各关节点更多的局部特征，使得解码器每层的输出既包含了手势的全局特征又包含了关节点的局部特征.根据图5，GUnet解码子网自身具有多尺度表达的特点，但现有姿态估计方法只利用了解码器的最后一层特征.因此，本文融合解码子网的各尺度特征图共同进行手部姿态估计.

由于GUnet结构中解码子网的各层特征图所表征的关节点数逐层增多，每个关节点的特征维度逐层减少，为了在尺度空间进行多尺度的特征融合表示，本文对解码子网每个尺度的特征图分别上采样到21个关节点，再通过一个图卷积层得到每个尺度的特征，最后将各个尺度的特征进行拼接得到手关节的多尺度特征表示.其计算过程如式(2)：

$$F_i=G_i^m(Unpooling_i(U_i)),\ i=5,4,3,2,1 \tag{2}$$

这里，U_i表示解码子网第i层的输出特征图，$Unpooling_i$表示反池化计算(即上采样)，$G_i^m(\cdot)$表示第i层的图卷积计算，F_i为第i尺度的特征图.解码子网的每一层都基于式(1)进行计算，从而得到同样大小的不同尺度的特征图，拼接这些不同尺度的特征图即得到张量表示的多尺度特征图.根据图5，手关节的多尺度特征图张量的计算如式(3).

$$\boldsymbol{\chi}=U_0\oplus F_1\oplus F_2\oplus F_3\oplus F_4\oplus F_5 \tag{3}$$

其中⊕表示特征融合，即特征拼接，$\boldsymbol{\chi}$表示多尺度融合后的张量.

鉴于手姿态的不同尺度特征图对最终的三维姿态预测的贡献不同，本文不是将多尺度的特征进行简单拼接后直接预测，而是引入注意力机制自学习每个尺度的特征图对手三维姿态估计的贡献，从而得到包含注意力信息的多尺度特征图.基于多尺度注意力机制的手姿态估计旨在充分挖掘各关节的更有表现力特征，从而提高手姿态估计的准确性.本文所提出的多尺度注意力模块如图6所示.多尺度注意力模块的输入是由改进的GUnet得到的S个尺度特征，本文将S个尺度的特征图组成特征张量$\boldsymbol{\chi}$，这里，$S=6$.为得到各尺度的权重，本文首先对特征张量$\boldsymbol{\chi}$进行全局平均池化操作，得到一维向量$\boldsymbol{z}\in\mathbb{R}^S$，计算公式如下：

$$z_s=\frac{1}{N\times K}\sum_{i=1}^{N}\sum_{j=1}^{K}\boldsymbol{\chi}_s(i,j) \tag{4}$$

这里，标量z_s表示第s尺度全局平均池化的结果，矩阵$\boldsymbol{\chi}_s\in\mathbb{R}^{N\times K}$表示第$s$尺度的特征图，$N$，$K$分别是特征图的高度和宽度，表示手关节的数量和每个关节的特征维度.$\boldsymbol{\chi}_s(i,j)$表示第s尺度特征图的第(i,j)个元素的特征值.

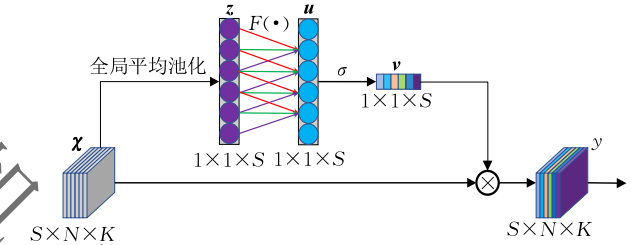


图6 多尺度注意力模块

对于全局平均池化得到的特征向量$\boldsymbol{z}$，如图6，本文使用一个一维卷积核对其进行卷积运算得到特征向量$\boldsymbol{u}$，并经非线性激活函数σ得到一个表征各尺度重要性的权重向量，计算如式(5)：

$$\boldsymbol{v}=\sigma(F(\boldsymbol{z})) \tag{5}$$

其中$\boldsymbol{v}$表示各尺度的权重向量；$F(\cdot)$表示一维卷积操作，一维卷积核的尺寸采用文献[24]的自适应选择策略；σ选用softmax函数.

将得到的权重向量$\boldsymbol{v}$的各分量分别作用到相应尺度的特征图，得到多尺度注意力模块的输出，即

$$y_s=F_s(\boldsymbol{\chi}_s,\boldsymbol{v}_s)=\boldsymbol{\chi}_s\boldsymbol{v}_s \tag{6}$$

如图5，MSA-GUnet网络的一个重要和基本的操作是图卷积，传统图卷积的计算如式(7)：

$$\boldsymbol{X}^{(l+1)}=\sigma(\boldsymbol{W}\boldsymbol{X}^{(l)}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\hat{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}) \tag{7}$$

其中，$\boldsymbol{X}^{(l)}\in\mathbb{R}^{K\times N}$表示第$l$层输入特征矩阵，$\boldsymbol{X}^{(l+1)}\in\mathbb{R}^{K'\times N}$表示第$l$层输出特征矩阵，$K$、$K'$分别表示第$l$层图卷积网络的每个节点的输入和输出特征维度，即每个手关节在第l、$l+1$层的特征维度.$\hat{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}$，$\boldsymbol{I}$为单位阵，$\boldsymbol{A}$表示邻接矩阵；$\tilde{\boldsymbol{D}}$表示$\hat{\boldsymbol{A}}$的度矩阵.$\boldsymbol{W}\in\mathbb{R}^{K'\times K}$表示权重矩阵，$\sigma$表示非线性激活函数.

上述传统图卷积网络假设所有图节点共享变换矩阵 $\boldsymbol{W}$，并且表示节点邻接关系的矩阵 $\hat{\boldsymbol{A}}\in[0,1]^{N\times N}$ 不能有效挖掘图结构中潜在的节点间关系，为此 Zhao 等人[25]提出语义图卷积，参考卷积神经网络的多卷积核特征表征能力，提出了学习通道级的边的权重，以更好挖掘节点之间的局部语义关系. 不过语义图卷积使用的是邻接矩阵 $\boldsymbol{A}$，而不是式(7)中的 $\hat{\boldsymbol{A}}$，缺少单位阵融合的邻接矩阵不能考虑节点自身的影响. 为了更好学习手部各关节之间的潜在的关联关系，本文改进了语义图卷积. 一方面，使用增加单位矩阵的 $\hat{\boldsymbol{A}}$，另一方面，定义了图残差掩膜 A^{res}. 改进的语义图卷积如式(8)：

$$\boldsymbol{X}^{(l+1)}=\Big\|_{d=1}^{D_{l+1}}\sigma(\vec{\boldsymbol{w}}_d\boldsymbol{X}^{(l)}\rho_i((\boldsymbol{M}_d\odot\hat{\boldsymbol{A}})+A_d^{\mathrm{res}})) \tag{8}$$

其中 ρ_i 是 softmax 操作，用于归一化矩阵元素；$\odot$ 表示矩阵元素级乘法运算：如果矩阵 $\hat{\boldsymbol{A}}$ 中元素 a_{ij} 的值为 1，那么返回矩阵 $\boldsymbol{M}_d$ 中 m_{ij} 的值，否则，返回值经过 ρ_i 操作得到近似 0 的值. 矩阵 $\hat{\boldsymbol{A}}$ 强制图中的节点 i 只计算自身以及与其相邻节点之间的权重；$\|$ 表示通道级联，$\vec{\boldsymbol{w}}_d$ 是变换矩阵 $\boldsymbol{W}$ 的第 d 行. A^{res} 初始化为近似零的随机值，并在网络训练过程中自动学习得到，旨在随机地、动态地增强或削弱邻接节点的连接权重，以更好学习节点之间的关系. σ 表示 ReLu 非线性激活函数.

在经典的 Unet[21] 网络中，编码器和解码器的每一层由两个卷积层叠加而成，以用更少的参数得到更大的感受野. 本文中 GUnet 网络的每一层仅由一个图卷积层构成，这是由于本文用图结构显式地表示手关节的连接关系，每个图节点表示手的一个关节，当图卷积层增多时，会过度平滑每个关节点，因而 GUnet 编解码器的每一层仅用一层上述改进的语义图卷积.

3.3　损失函数

深度学习网络的本质是求解一个非线性优化问题，优化的目标通常被形式化为一个复杂函数，并称其为损失函数；网络训练的过程就是优化损失函数的过程，所以一个好的损失函数是重要的. 本文的目标是精准预测输入图像中手的三维关节位置，所以希望预测的手三维关节位置和标注的手三维关节位置之间的误差尽可能小.

如图 1，基于两阶段法的手三维姿态估计包括二维姿态估计和三维姿态估计两部分，所以总体损失函数包括二维姿态估计损失和三维姿态估计损失；同时，二维姿态估计损失源于两个环节，即初始估计损失和精细化估计损失，因此，网络的总体损失函数定义为

$$L=\lambda_{\mathrm{init}}L_{\mathrm{init}}+\lambda_{2\mathrm{D}}L_{2\mathrm{D}}+\lambda_{3\mathrm{D}}L_{3\mathrm{D}} \tag{9}$$

其中 $\lambda_{\mathrm{init}}=0.01$、$\lambda_{2\mathrm{D}}=0.01$、$\lambda_{3\mathrm{D}}=1$，超参的取值源于基线方法[20]. L_{init} 为初始的二维姿态估计损失，$L_{2\mathrm{D}}$、$L_{3\mathrm{D}}$ 分别为精细化的二维姿态估计损失和三维姿态估计损失.

初始的二维姿态损失定义为

$$L_{\mathrm{init}}=\sum_{j=1}^{N}\|\boldsymbol{h}_j-\hat{\boldsymbol{h}}_j\|_2^2 \tag{10}$$

其中 $\boldsymbol{h}_j$ 表示标注的二维姿态的第 j 个关节的二维坐标，$\hat{\boldsymbol{h}}_j$ 表示预测的初始二维姿态的第 j 个关节的二维坐标，N 表示手关节的总数.

为了实现精准的二维姿态估计，本文设计了物理和对称双分支网络，因此精准的二维姿态损失为两个分支的和，其损失函数如下：

$$L_{2\mathrm{D}}=\sum_{m=1}^{2}\sum_{j=1}^{N}\|\boldsymbol{h}_j^m-\hat{\boldsymbol{h}}_j^m\|_2^2 \tag{11}$$

其中，m 表示分支标号，$m=1$ 表示物理连接的分支；$m=2$ 表示对称连接的分支. N 表示手的总关节数. $\boldsymbol{h}_j^m$ 表示标注的第 m 个分支第 j 个关节的二维姿态，$\hat{\boldsymbol{h}}_j^m$ 表示预测的第 m 个分支第 j 个关节二维姿态坐标.

三维姿态估计的损失 $L_{3\mathrm{D}}$ 为

$$L_{3\mathrm{D}}=\lambda_{\mathrm{pose}}L_{\mathrm{pose}}+\lambda_{\mathrm{len}}L_{\mathrm{len}}+\lambda_{\mathrm{dir}}L_{\mathrm{dir}} \tag{12}$$

其中 λ_{pose}、λ_{len} 和 λ_{dir} 分别为三维姿态损失 L_{pose}、骨骼长度损失 L_{len} 和骨骼方向损失 L_{dir} 的权重超参数，$\lambda_{\mathrm{pose}}=1$、$\lambda_{\mathrm{len}}=0.01$、$\lambda_{\mathrm{dir}}=0.1$.

姿态损失 L_{pose} 定义为

$$L_{\mathrm{pose}}=\sum_{m=1}^{2}\sum_{j=1}^{N}\|\boldsymbol{\phi}_j-\hat{\boldsymbol{\phi}}_j^m\|_2^2 \tag{13}$$

其中 $\boldsymbol{\phi}_j$ 表示标注的第 j 个关节的三维坐标，$\hat{\boldsymbol{\phi}}_j^m$ 表示第 m 个分支预测的第 j 个关节三维坐标.

骨骼长度损失 L_{len} 和骨骼方向损失 L_{dir} 如下：

$$L_{\mathrm{len}}^m=\sum_{i,j}\Big|\|\boldsymbol{b}_{i,j}\|_2-\|\hat{\boldsymbol{b}}_{i,j}^m\|_2\Big| \tag{14}$$

$$L_{\mathrm{dir}}^m=\sum_{i,j}\left\|\frac{\boldsymbol{b}_{i,j}}{\|\boldsymbol{b}_{i,j}\|_2}-\frac{\hat{\boldsymbol{b}}_{i,j}^m}{\|\hat{\boldsymbol{b}}_{i,j}^m\|_2}\right\|_2 \tag{15}$$

其中 $\boldsymbol{b}_{i,j}$ 表示依据标注坐标计算的手部第 i 个关节和第 j 个关节之间的骨骼矢量，$\hat{\boldsymbol{b}}_{i,j}^m$ 表示基于第 m 分支预测的手部第 i 个关节和第 j 个关节之间的骨骼矢量，即

$$\boldsymbol{b}_{i,j}=\boldsymbol{\phi}_i-\boldsymbol{\phi}_j \tag{16}$$

$$\hat{\boldsymbol{b}}_{i,j}^m=\hat{\boldsymbol{\phi}}_i^m-\hat{\boldsymbol{\phi}}_j^m \tag{17}$$

其中 $\boldsymbol{\phi}_i$ 和 $\boldsymbol{\phi}_j$ 表示标注的第 i 和第 j 个关节点的坐标，$\hat{\boldsymbol{\phi}}_i^m$ 和 $\hat{\boldsymbol{\phi}}_j^m$ 表示预测的第 m 分支第 i 和第 j 个关节

点的坐标.每个分支按照式(14)、(15)计算各自分支的骨骼长度损失和骨骼方向损失,然后将其求和:

$$L_{\text{len}}=\sum_{m=1}^{2}L_{\text{len}}^{m} \tag{18}$$

$$L_{\text{dir}}=\sum_{m=1}^{2}L_{\text{dir}}^{m} \tag{19}$$

综合式(13)、(18)和式(19)的结果即求得式(12),进一步地综合式(10)、(11)和式(12)即可得式(9)总体损失函数的表达式.

4 实 验

为了验证本文所提出的姿态估计方法的有效性,本文设计了三个实验,前两个实验是关于手姿态估计性能的定量评估,第三个实验是手姿态估计结果的可视化,用于主观评测所提方法的性能.定量评估实验主要包括消融实验:所提双分支框架的有效性验证、多尺度注意力实验验证、改进语义图卷积的有效性验证,与现有方法的对比实验以及预训练并微调的评估实验.本文在公开的数据集上进行了上述实验,实验的软硬件配置信息如表1所示.

表1 实验软硬件的配置信息

项目	配置
CPU	Intel(R) Core(TM) i9-10900X
内存	128GB
GPU	NVIDIA GeForce GTX 3090
操作系统	Ubuntu18.04 64位
Cuda	cuda11.0.221 with cudnn 8.0.3
语言环境	Python3.7

4.1 数据集

本文基于 Obman、STB、FreiHand 三个数据集进行实验,其中 Obman 合成数据集用于网络预训练,STB 和 FreiHand 两个真实数据集用于微调和测试,所有实验数据都标注了 21 个关节点的三维坐标.

(1) Obman 数据集[26].该数据集中的手势图像是基于渲染手网格模型得到的合成数据集.包含141550个训练数据、6463个验证数据和6285个测试数据,主要用于网络预训练.

(2) STB 数据集[27].手姿态估计研究的经典数据集之一.该数据集是在六种不同环境下采集的,包括两种不同的手势,一种是简单的数字手势,另一种是复杂的随机手势.数据集共包括12个视频序列,每个序列由1500帧手势图像组成,总计18000张图像,数据集包含了手掌中心和手关节的3D标注.本文参考文献[4,16,28]的做法,选择15000张用于训练,3000张用于测试.实验前将所有图像裁剪到分辨率为224×224的图像,由于该数据集中的手势图像源于手势动作视频,因此相邻帧图像的相似度大,为此,采取间隔采样的方法构建测试集,其它样本构成训练集.实验时,由于该数据集标注的是掌关节,而21个关节手模型标注的是腕关节,所以按照参考文献[4]的做法将掌关节转换为腕关节.

(3) FreiHand 数据集[29].该数据集是2019年发布的,包括32个对象的手势.手势数据背景多样、视点多变、动作复杂,一些手势是与物体交互的动作.本文选取背景最复杂的33000张图像进行实验.不同于STB数据集,该数据集中的图像是杂乱随机的.本文随机选取80%的图像数据用于训练,其余20%用于测试.

4.2 评估指标

为评价手三维姿态估计的准确性,本文采用该领域最常用的平均关节误差指标进行度量,其描述为标注的三维关节位置与预测的三维关节位置之间的欧氏距离,计算公式如下:

$$MPJPE=\frac{1}{T}\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}\left\|\boldsymbol{J}_i^{(t)}-\hat{\boldsymbol{J}}_i^{(t)}\right\|_2 \tag{20}$$

其中 $MPJPE$ 是平均关节误差;T 表示样本数量,N 表示关节点的数量;$\boldsymbol{J}_i^{(t)}$ 表示第 t 张图像的标注的第 i 个关节的三维位置,$\hat{\boldsymbol{J}}_i^{(t)}$ 表示第 t 张图像的预测的第 i 个关节的三维位置.

4.3 网络参数设置

本文实验在开源的网络框架 Pytorch 上进行.首先使用合成数据集 Obman 对网络模型进行预训练,然后在真实数据集上微调.预训练的初始学习率设置为0.001,每500轮学习率乘以0.1,共训练5000轮.真实数据集上微调的设置如下:使用真实数据集将预训练模型端到端训练1000轮,初始学习率为0.001,每100轮学习率乘以0.9.

4.4 实验结果

4.4.1 消融实验

本文分别在 STB 数据集和 FreiHand 数据集上进行了消融实验,以验证本文所提出的双分支、多尺度特征表示、多尺度注意力以及改进的语义图卷积的有效性.本节最后也对预训练和微调的作用进行了实验评估.

为了验证本文提出的双分支网络框架的有效性,本文在表2比较了基线方法(Baseline[20])、基于物理连接的单分支结构(Physical Branch,PB)、基于对称连接的单分支结构(Symmetrical Branch,SB)以及基于物理和对称双分支(Double Branch,DB)互补结构预测手关节三维位置的平均关节误差.关于双分支结构的融合,本文给出了基于均值融

合和基于学习的权重加权融合两种策略，为便于区分，将基于平均融合策略的方法称为 DB，而将基于学习权重融合的方法称为 DBW.

表 2　双分支有效性验证实验　（单位：mm）

方法	STB 数据集	FreiHand 数据集
Baseline[20]	8.327	9.006
PB	8.852	9.888
SB	10.078	9.194
DB	8.044	**8.637**
DBW	**7.986**	8.646
DBMSA-SGCN	5.725	**8.158**
DBWMSA-SGCN	**5.652**	8.261

根据表 2，双分支结构优于基线方法，单纯物理分支或对称分支的预测结果比基线方法差. 这是由于基线方法中关节之间的邻接关系是学习得到的，而本文中的两个分支分别预先指定了关节之间的物理和对称连接关系，这种特定的连接关系是有局限性的，通常只对部分手势更有表达力，因此，单纯基于一种限定的手关节连接关系预测手的三维关节位置不如基于自适应学习连接关系方法的预测结果；但是，当使用具有互补连接关系的两种结构共同建模手关节时，得到的平均关节误差比基线方法小，表明手关节的物理连接和对称连接可以互补建模手关节的复杂连接关系.

为了测试两个分支的融合策略，本文分别在提出的基本 DB 框架以及 DBMSA-SGCN 上进行了实验. 根据表 2 的后四行不难看出，融合策略与数据集有关. 在 STB 数据集上，基于学习权重的融合策略更好，而在 FreiHand 数据集上，基于均值的融合策略更优. 这是由于 STB 数据集相对简单，基于学习权重的加权融合方法能够从数据中学出每个关节在两种结构表征中的相对重要性，所以略优于均值融合策略，而 FreiHand 数据集的数据复杂，难以挖掘两种拓扑结构表达能力的统计特性，所以均值融合策略更佳. 根据表 2，基于学习权重的融合策略并不显著提升姿态估计准确性，所以本文后续实验均采用了平均融合的策略.

为进一步验证本文所提双分支框架的互补性和必要性，本文选取了两张代表性手势图像，分别报告单纯通过物理连接或对称连接进行姿态估计的主客观结果. 主观的可视化结果见第 3.1 节图 4. 根据图 4 不难推断，第一行的手势适合用对称分支建模而第二行的手势适合用物理分支建模. 本文也给出了这两个手势的各手指的平均关节误差，实验结果如表 3 所示. 这里 PB1 和 SB1 表示第一行的手势的单分支实验结果；PB2 和 SB2 表示第二行的手势的单分支实验结果. 根据表 3 展示的客观结果，可得出与图 4 主观结果同样的结论，对于第一行手势，对称分支有更好的预测结果；而对于第二行手势，物理分支有更好的预测结果. 综上，图 4 和表 3 进一步说明了两个不同分支的互补性和必要性.

表 3　不同分支上的手指平均关节误差（单位：mm）

	拇指	食指	中指	无名指	小指
PB1	**2.792**	4.816	6.662	**8.426**	7.902
SB1	3.407	**2.004**	**5.381**	8.829	**5.759**
PB2	5.870	**2.641**	**2.432**	**2.887**	**2.643**
SB2	**3.848**	3.269	4.109	3.014	3.743

为了验证本文提出的单分支多尺度注意力模块的有效性，本文首先在物理分支上分别测试了增加多尺度表示和多尺度注意力模块的客观性能指标，然后给出了在双分支网络框架的每一分支分别增加多尺度表示和多尺度注意力模块后的实验结果. 表 4 展示了单物理分支（PB）、物理分支多尺度表示（PBMS）、物理分支多尺度注意力（PBMSA）、双分支（DB）、双分支多尺度表示（DBMS）以及双分支多尺度注意力模块（DBMSA）的实验结果. 根据表 4 可以看出，多尺度表示和多尺度注意力模块对单分支和双分支两种结构的预测都有很大的性能提升，其中多尺度表示对手姿态估计的预测性能提升更大，如 PBMS 相对 PB 在 STB 和 FreiHand 数据集上的预测误差分别降低了 1.590 mm 和 1.277 mm；DBMS 相对 DB 在 STB 数据集上的平均关节误差降低了 1.959 mm，说明多尺度表示能表达各关节不同的上下文特征，因而更有利于手的三维姿态估计. 根据实验结果不难看出，多尺度表示在 STB 数据集上的提升作用明显，可能的原因是 STB 数据集的图像分辨率较高，所以进行多尺度表示后取得了更好的效果，而 FreiHand 数据集的图像分辨率比较低，所以多尺度表示的作用提升不及高分辨的 STB 数据集明显.

表 4　多尺度注意力有效性验证实验（单位：mm）

方法	STB 数据集	FreiHand 数据集
PB	8.852	9.888
PBMS	7.262	8.611
PBMSA	7.018	8.492
DB	8.044	8.637
DBMS	6.085	8.401
DBMSA	**5.898**	**8.362**

本文也以物理分支的可视化结果为例，定性展示了多尺度表示和多尺度注意力模块对本文手三维姿态估计的作用. 如图 7，“握拳”的手势通过物理分支预测后的结果不理想，而在增加多尺度表示和多尺度注意力模块后，物理分支的预测结果有显著提

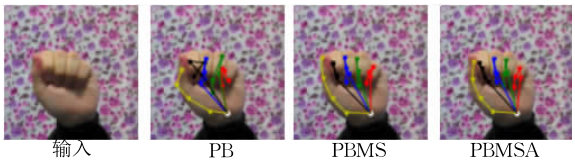


图 7 单分支多尺度注意力消融实验的可视化结果

升,表明本文提出的多尺度注意力模块有助于改善单分支的预测性能.

为了验证本文提出的改进语义图卷积的有效性,本文分别用所提语义图卷积替换基线方法和 DBMSA 中的图卷积,实验结果如表 5 所示. 根据表 5,改进图卷积模块后,Baseline-SGCN 和 DBMSA-SGCN 相对 Baseline 和 DBMSA 的平均关节误差均降低,不难看出使用改进的语义图卷积后,能够提取节点之间更准确的语义信息,从而验证了所提语义图卷积的有效性.

表 5 语义图卷积有效性验证实验 (单位:mm)

方法	STB 数据集	FreiHand 数据集
Baseline[20]	8.327	9.006
Baseline-SGCN	7.256	8.406
DBMSA	5.898	8.362
DBMSA-SGCN	**5.725**	**8.158**

本文也给出了不同误差阈值下的正确关节点比例(Percentage of Correct Keypoints,PCK)曲线,STB 和 FreiHand 数据集上的 PCK 曲线如图 8 和图 9 所示. 根据实验结果,本文所提方法优于基线方法,并且增加不同模块后性能不断提升. 本文中 DB、DBMS、DBMSA 和 DBMSA-SGCN 的 PCK 曲线下面积(the Area Under the Curve,*AUC*)依次增大,在 STB 数据上,DBMSA-SGCN 的 *AUC* 达到 0.893,在 FreiHand 数据集上的 *AUC* 达到 0.834,均为最高. 虽然一些文献在 STB 数据集的 *AUC* 高达 0.997[30],不过它们的计算是基于预测误差在 20 mm 以上的

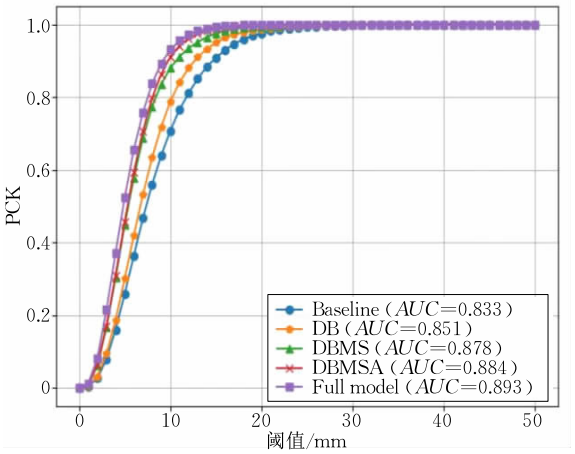


图 8 STB 数据集上的 PCK 曲线

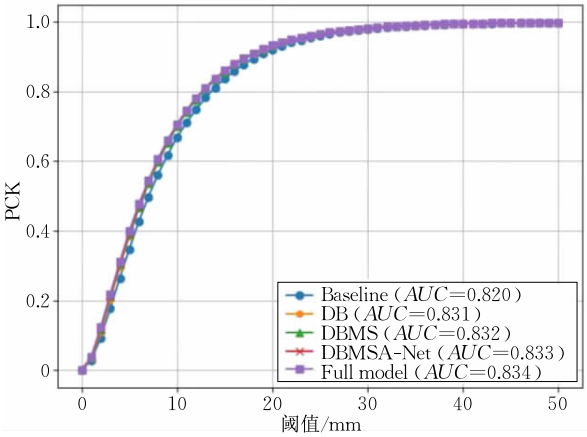


图 9 FreiHand 数据集上的 PCK 曲线

阈值计算. 如果误差阈值起点同样为 20 mm,本文方法在 STB 数据集上的 *AUC* 可达 0.999. 该实验进一步表明了所提模型的有效性.

为了评估本文基于 Obman 合成数据集进行预训练并在真实数据集微调的作用,本文设计了如下两个实验:只用预训练网络而不微调,直接在测试集上进行测试;不进行预训练,直接在 STB 和 FreiHand 数据集进行训练并测试,实验结果如表 6. 根据表 6,尽管预训练数据集很大,但由于合成数据集和真实数据集的分布差异,只用预训练模型而不微调,会产生很大的估计误差;而预训练并微调相对直接在真实数据集训练能够减少 0.3 mm 左右的平均关节误差,这是由于合成数据集有精准的手部关节的三维坐标标注,因而有助于学习到图像到手三维关节之间的非线性映射关系. 综上,本文基于合成数据集进行预训练并在真实数据集进行网络参数微调是必要的.

表 6 预训练并微调对平均关节误差的影响评估 (单位:mm)

DBMSA-SGCN	STB 数据集	FreiHand 数据集
预训练而不微调	23.311	35.619
没有预训练模型	6.061	8.430
预训练并微调	**5.725**	**8.158**

4.4.2 与现有方法的对比实验

本文也与一些基于单张彩色图像估计手三维姿态的现有方法进行了比较,对比方法的选择依据是近年来在 STB 和 FreiHand 数据集上实验的方法. 实验结果如表 7 和表 8 所示.

表 7 STB 数据集与现有方法的比较

方法	平均关节误差/mm
Theodoridis 等人[30]	6.930
Spurr 等人[28]	8.560
Ge 等人[16]	6.370
Yang 等人[1]	7.050
DBMSA	**5.898**
DBMSA-SGCN	**5.725**

表 8　FreiHand 数据集与现有方法的比较

方法	平均关节误差/mm
Parelli 等人[31]	11.000
Doosti 等人[20]	9.006
Ge 等人[16]	15.360
DBMSA	**8.362**
DBMSA-SGCN	**8.158**

根据表 7 可知，本文方法在 STB 数据集上比 Theodoridis 等人方法[30] 的平均关节误差降低 1.205 mm，比 Ge 等人[16] 方法降低 0.645 mm. 根据表 8 可见，在 FreiHand 数据集上比 Parelli 等人提出方法[31] 的平均关节误差减小 2.842 mm，比 Doosti 等人方法[20] 降低 0.848 mm.

4.4.3　定性结果

本文给出了所提方法在 STB 数据集和 FreiHand 数据集上的一些可视化结果，以直观展示所提方法的有效性. 可视化方法是将预测得到的三维坐标投影到输入的二维手势图像.

如图 10 和图 11 所示，相比单分支的 PB 和 SB 网络结构，本文所提出的 DB、DBMSA 以及 DBMSA-SGCN 具有更好的三维手姿态预测结果. 其中，DBMSA-SGCN 网络的姿态估计结果最好，与 Ground truth 最接近. 如图 10，基于 DBMSA-SGCN 模型预测的两个手势的姿态和 Ground truth 接近一致；图 11 中，基于 DBMSA-SGCN 模型预测的小指的可视化结果优于 Ground truth. 如图 11 第一行所示，基于 DBMSA-SGCN 预测的结果相对 Ground truth 有更短的小指，这是更合理的，说明本文利用手指关节间的结构约束是有效的；再如图 11 第二行所示，相对于 Ground truth，基于 DBMSA-SGCN 能预测出小指的弯曲结构，这显然是更合理的. 根据图 10 和图 11 也不难看出，基于 DB 的预测结果优于单一的 PB 或 SB 的预测结果，特别地，图 11 中两个手势的 DB 预测结果明显优于单独一个分支的结果. 进一步地，引入多尺度注意力后，基于 DBMSA 的结果优于 DB，图 11 第一行的可视化结果可明显地验证这一点. 增加语义图卷积模块后的 DBMSA-SGCN 结果优于 DBMSA 结果，不难看出，在图 10 和图 11 列出的实验结果中，DBMSA-SGCN 的结果最好，说明增加改进语义图卷积后提取的关节特征更好表征了关节之间的语义关系.

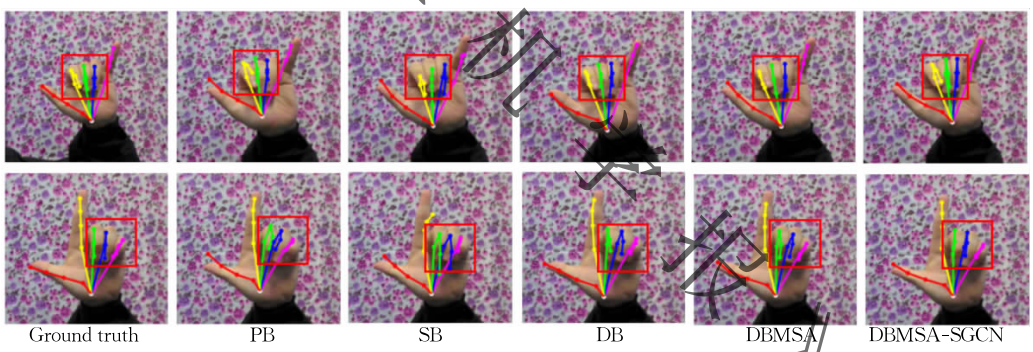


图 10　STB 数据集的部分可视化结果

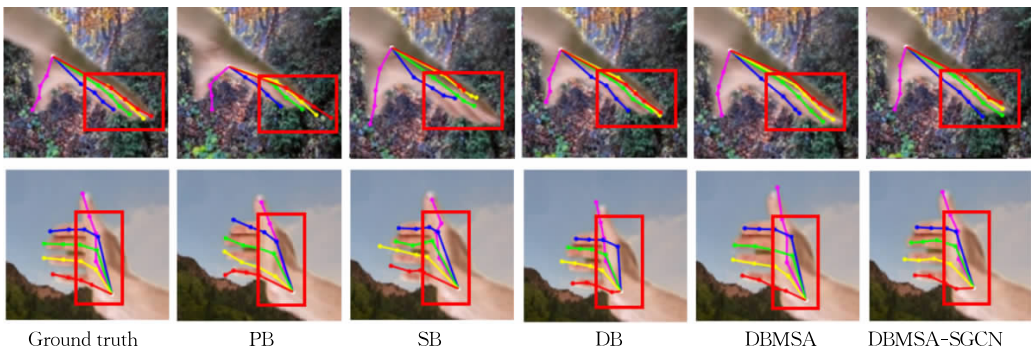


图 11　FreiHand 数据集的部分可视化结果

4.4.4　模型的参数量和计算复杂度

本文给出了所提模型的参数量和计算复杂度分析，如表 9. 由于所提网络模型的参数量和运算复杂度大部分消耗在图像特征提取阶段的 ResNet50 网络，所以本文给出了 ResNet50 模块、Baseline 以及本文所提的 DBMSA-SGCN 的参数量和计算复杂度. 每张测试图像在 3.70 GHz CPU 和 NVIDIA RTX3090 上的姿态估计时间为 7.8 ms.

表 9 主要模型的参数量和计算复杂度比较

网络	模型参数量/M	每秒浮点运算次数/G
ResNet50	23.50	8.21
Baseline	23.87	8.28
DBMSA-SGCN	24.72	10.76

5 结 论

本文提出了一种面向单张 RGB 图像的基于双分支结构的手三维姿态估计框架，其中，物理分支建模手的同一手指不同关节之间的固有连接关系，而对称分支建模手的不同手指同一关节之间的对称运动属性. 手的两种不同的图结构描述互补约束有助于更好地建模手关节之间的关联关系，解决了手关节复杂难建模问题，降低了手姿态估计误差，从而提升了手姿态估计的准确性；进一步地，在每一分支，提出了基于多尺度注意力 GUnet 和改进语义图卷积的手姿态估计方法，多尺度表示、多尺度注意力以及改进的语义图卷积均有助于提升手三维姿态估计的准确性，其中多尺度表示有助于显著提升单分支手姿态估计的准确性. 本文在两个公开的数据集上分别进行了消融实验，并与现有方法进行了对比，实验结果验证了本文所提方法的有效性.

未来的工作一方面将考虑基于超图神经网络建模手关节的复杂关联关系；另一方面，鉴于手的三维姿态标注困难，将考虑基于弱监督、自监督以及领域适配等方法进行手姿态估计.

参 考 文 献

[1] Yang L, Li S, Lee D, et al. Aligning latent spaces for 3D hand pose estimation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, South Korea, 2019: 2335-2343

[2] Tompson J J, Jain A, LeCun Y, Bregler C. Joint training of a convolutional network and a graphical model for human pose estimation//Proceedings of the Conference on Neural Information Processing Systems. Montreal, Canada, 2014: 1799-1807

[3] Bogo F, Kanazawa A, Lassner C, et al. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image//Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands, 2016: 561-578

[4] Zimmermann C, Brox T. Learning to estimate 3D hand pose from single RGB images//Proceedings of the IEEE/CVF International Conference on Computer Vision. Hawaii, USA, 2017: 4903-4911

[5] Guo S, Rigall E, Qi L, et al. Graph-based CNNs with self-supervised module for 3D hand pose estimation from Monocular RGB. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(4): 1514-1525

[6] Jiang X, Ma X. Dynamic graph CNN with attention module for 3D hand pose estimation//Proceedings of the International Symposium on Neural Networks. Springer, Cham, 2019: 87-96

[7] Cai Y, Ge L, Liu J, et al. Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks //Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, South Korea, 2019: 2272-2281

[8] Gao H, Ji S. Graph U-nets//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 2083-2092

[9] Wang J, Yan S, Xiong Y, et al. Motion guided 3D pose estimation from videos//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 764-780

[10] Tompson J, Stein M, Lecun Y, et al. Real-time continuous pose recovery of human hands using convolutional networks. ACM Transactions on Graphics, 2014, 33(5): 1-10

[11] Wan C, Probst T. Dense 3D regression for hand pose estimation //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake, USA, 2018: 5147-5156

[12] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation//Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands, 2016: 483-499

[13] Du K, Lin X, Sun Y, et al. CrossInfoNet: Multi-task information sharing based hand pose estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 9896-9905

[14] 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//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake, USA, 2018: 5079-5088

[15] Baek S, Kim K I, Kim T K. Pushing the envelope for RGB-based dense 3D hand pose estimation via neural rendering// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 1067-1076

[16] Ge L, Ren Z, Li Y, et al. 3D hand shape and pose estimation from a single RGB image//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 10833-10842

[17] Cai Y, Ge L, Cai J, Magnenat-Thalmann N, Yuan J. 3D hand pose estimation using synthetic data and weakly labeled RGB images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(11): 3739-3753

[18] Hu T, Wang W, Lu T. Hand pose estimation with attention-and-sequence network//Proceedings of the Pacific Rim Conference on Multimedia. Hefei, China, 2018: 556-566

[19] Kazakos E, Nikou C. Kakadiaris I. On the fusion of RGB and depth information for hand pose estimation//Proceedings of the International Conference on Image Processing. Athens, Greece, 2018: 868-872

[20] Doosti B, Naha S, Mirbagheri M, et al. Hope-net: A graph-based model for hand-object pose estimation//Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 6608-6617

[21] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation//Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany, 2015: 234-241

[22] Chen L C, Yang Y, Wang J, et al. Attention to scale: Scale-aware semantic image segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 3640-3649

[23] Lowe D G. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004, 60(2): 91-110

[24] Wang Q, Wu B, Zhu P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 11531-11539

[25] Zhao L, Peng X, Tian Y, et al. Semantic graph convolutional networks for 3D human pose regression//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 3425-3435

[26] Hasson Y, Varol G, Tzionas D, et al. Learning joint reconstruction of hands and manipulated objects//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 11807-11816

[27] Zhang J, Jiao J, Chen M, et al. A hand pose tracking benchmark from stereo matching//Proceedings of the IEEE International Conference on Image Processing. Beijing, China, 2017: 982-986

[28] Spurr A, Song J, Park S, et al. Cross-modal deep variational hand pose estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake, USA, 2018: 89-98

[29] Zimmermann C, Ceylan D, Yang J, et al. FreiHand: A dataset for markerless capture of hand pose and shape from single RGB images//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, South Korea, 2019: 813-822

[30] Theodoridis T, Chatzis T, Solachidis V, et al. Cross-modal variational alignment of latent spaces//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA, 2020: 960-961

[31] Parelli M, Papadimitriou K, Potamianos G, et al. Exploiting 3D hand pose estimation in deep learning-based sign language recognition from RGB videos//Proceedings of the European Conference on Computer Vision Workshop. Glasgow, UK, 2020: 249-263

MA Sheng-Lei, M. S. His research interests include image processing and machine learning.

LI Jing-Hua, Ph. D., associate professor, M. S. supervisor. Her research interests include image processing and machine learning.

KONG De-Hui, Ph. D., professor, Ph. D. supervisor. Her research interests include computer graphics, computer vision, virtual reality.

WANG Li-Chun, Ph. D., professor, Ph. D. supervisor. Her research interests include artificial intelligence, human-computer interaction.

WANG Shao-Fan, Ph. D., associate professor, M. S. supervisor. His research interests include pattern recognition, machine learning.

YIN Bao-Cai, Ph. D., professor, Ph. D. supervisor. His research interests include multimedia technology, cross-media intelligence, video coding.

Background

3D hand pose estimation aims to predict the location of 3D hand joints from the one or multiple images, which is one of the hot topics on computer vision. 3D hand pose estimation has wide applications in the fields of robotics, human-computer interaction, virtual reality and augmented reality etc.

3D hand pose estimation from single RGB image is a very promising but challenging research. The main difficulty lies in self-occlusion, depth ambiguity and insufficient 3D annotation. Many novel deep learning methods such as CNN, VAE, GAN, Pointnet, GCN and their improved versions, have been applied to increase the accuracy of 3D hand pose estimation from single RGB image.

This paper applied the latest popular GCN and GU-net to 3D hand pose estimation based on single RGB, especially, made full use of the structure prior of hand joints, which explicitly models the relationship of the hand joints existing inside and between fingers. Furthermore, multi-scale representation and scales attention is integrated into the proposed network architecture to improve the accuracy of 2D to 3D hand pose regression. Experimental results on two public datasets verified the effectiveness of the proposed method and the mean joint error decreased 0.6 mm and 0.8 mm, respectively.

This work is supported by the National Natural Science Foundation of China (Nos. 62172022, U21B2038, 61876012) and the Natural Science Foundation of Beijing (No. 4202003).