文章编号　1004-924X(2025)20-3265-16

联合多模态特征与结构感知的手物交互姿态估计

王文润[1,2]，党建武[1,2*]，王阳萍[2,3]，任鹏百[1,3]，潘　瑞[2]

(1. 兰州交通大学 轨道交通信息与控制国家级虚拟仿真实验教学中心，
甘肃 兰州 730070；
2. 兰州交通大学 电子与信息工程学院，甘肃 兰州 730070；
3. 甘肃省人工智能与图形图像处理工程研究中心，甘肃 兰州 730070)

摘要：现实世界中手不可避免地要与物体进行交互，因此理解人手与物体的交互行为与意图具有重要的研究意义。本文针对手与物体交互过程中的相互遮挡、手部自遮挡及复杂交互背景等因素导致姿态估计精度低的问题，提出一种联合多模态特征与结构感知的手部与交互物体三维姿态估计方法。该方法利用彩色图像和深度图像的多模态特征实现信息互补，有效解决背景复杂、手部自遮挡及手物相互遮挡的问题；其次，基于图结构分别设计手部、交互物体及手物交互结构感知模块，辅助估计更加合理和准确的手与交互物体的二维姿态；最后，将获取的二维姿态与深度图像中的深度信息进行合并，再利用纹理特征对合并得到的三维姿态进一步优化得到最终的手物交互三维姿态。为了验证本文方法的有效性，在 FPHA，HO-3D 等数据集开展了系列实验，手部和交互物体的姿态误差分别降低到 9.62 mm 和 14.37 mm。实验结果表明，所提方法优于现有的手物交互姿态估计方法，具有较强的鲁棒性和泛化性。

关　键　词：手物姿态估计；图卷积网络；多模态特征；结构感知

中图分类号：TP391.41　　**文献标识码**：A

doi：10.37188/OPE.20253320.3265　　**CSTR**：32169.14.OPE.20253320.3265

Hand-object interaction pose estimation integrating multi-modal features and structure awareness

WANG Wenrun[1,2], DANG Jianwu[1,2*], WANG Yangping[2,3], REN Pengbai[1,3], PAN Rui[2]

(1. *National Virtual Simulation Experimental Teaching Center for Rail Transit Information and Control, Lanzhou Jiaotong University, Lanzhou* 730070, *China*;
2. *School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou* 730070, *China*;
3. *Gansu Provincial Engineering Research Center for Artificial Intelligence and Graphic & Image Processing, Lanzhou* 730070, *China*)
* *Corresponding author, E-mail: dangjw@mail.lzjtu.cn*

Abstract: In the real world, hands inevitably interact with objects. Understanding the interaction behaviors and intentions between human hands and objects is of great research significance. This paper tackled the low-accuracy pose-estimation issue during hand-object interaction, caused by mutual hand-object occlu-

收稿日期：2025-07-23；**修订日期**：2025-08-27.

基金项目：国家自然科学基金项目(No. 62067006；No. 62367005)；学校青年基金科学项目(No. 2022012)

sion, hand self-occlusion, and complex backgrounds. A 3D pose-estimation method for hands and interacting objects, which combined multi-modal features and structure awareness, was proposed. This method exploited the multi-modal features of color and depth images for information complementarity, effectively addressing complex backgrounds, hand self-occlusion, and hand-object mutual occlusion. Second, graph-structure-based awareness modules for the hand, the object, and their interaction were designed to help estimate more reasonable and accurate 2D poses. Finally, the obtained 2D poses were merged with depth-image depth information, and texture features were used to optimize the merged 3D poses for the final hand-object interaction 3D pose. To verify the method's effectiveness, experiments were conducted on datasets like FPHA and HO-3D. The hand and object pose errors are reduced to 9.62 mm and 14.37 mm, respectively. Results show the proposed method outperforms existing ones and has strong robustness and generalization.

Key words: hand-object pose estimation; graph convolutional network; multi-modal features; structure awareness

1 引　言

在人类日常生活中,手部作为关键的交互工具,其灵活程度和姿态变化多样性是人体其他部位所不能替代的。现有的手部姿态估计可分为借助外接设备的有标记方法和基于视觉的无标记方法[1]。然而,基于视觉的无标记方法以其自然交互体验的优势,使得用户体验效果更佳,也驱使研究者们不断探索基于纯视觉的高精度手部姿态估计技术。目前,针对空手条件下的手部姿态估计已达到很高的精度,估计误差可降至5 mm以内[2],能可靠支撑虚拟交互、手语识别等任务;但当手与被操作物体进行交互时,因严重遮挡和复杂外观,导致模型性能急剧下降,其效果并不令人满意,估计误差通常超过10 mm[3]。而在现实生活中,人类依赖双手抓取物体、操作器械等,因此对交互场景中手部姿态的精确估计,并协同估计操作物体姿态,对人手动作行为特征的理解,以及指导虚拟手和仿生机械手实现精细化的手物交互动作模拟都有重要的研究意义[4-5],在增强现实和虚拟现实[6-8]、人机交互[9-12]、机器人操作[13]、机器人辅助手术[14]以及具身人工智能[15-17]等各个领域都有广泛的应用。然而,由于相互遮挡、外观模糊等带来的困难,如何实现高精度的手物交互姿态估计,仍是当前研究亟待攻克的核心难题。

近年来,随着深度学习、彩色图像与深度图像(Red Green Blue-Depth, RGB-D)传感器的迅速发展,空手状态下的手部姿态估计取得了显著进展。按照输入图像的类型进行划分,主要包括基于深度图像[18-22]和基于彩色图像(Red Green Blue, RGB)[23-29]两种。基于深度图像的方法因包含丰富的三维空间信息而在姿态估计精度上表现优异,但由于现阶段的深度传感器成像范围受限和分辨率不高的问题,其应用主要局限于一些科学实验和娱乐性游戏等场景[1]。2017年,Goudie等人[18]首次提出从深度图像中采用端到端的两阶段卷积神经网络方法来估计三维(Three Dimensions, 3D)手部姿态。第一阶段将图像中的像素分为背景、手部和物体部分;第二阶段,将分割后的图像与输入的深度图像进行拼接后输入到网络中获取21个手部关节的3D位置。基于RGB图像的手部姿态估计方法因包含丰富的纹理特征、高分辨率及应用范围广而受到学者的广泛关注,但由于深度信息缺失会导致不可靠的深度估计[30],因此对手部3D姿态的估计也更具挑战性。2017年,Zimmermann等人[23]首次设计两阶段网络架构直接从单张RGB图像预测手部关节姿态。第一阶段使用HandSegNet和PoseNet网络对图像分别进行分割和二维(Two Dimensions, 2D)姿态估计;第二阶段使用Pose-Prior和Viewpoint分别估计3D姿态和视点。通过查阅文献总结可知,这种"先2D后3D"的分阶段手部姿态估计方法,凭借其更高的精度优势,获得了学术界的广泛认可。当前基于深度学习的手部姿态估计方法普遍是在两段式的基础上

引入多尺度特征融合和注意力机制[31-34]、手部先验结构[35-38]等知识对卷积神经网络[39]、图卷积神经网络[40]及Transformer神经网络[41]等不同网络不断优化迭代，从而提高手部姿态估计的精度。然而，上述研究工作仅专注于空手状态下的手部3D姿态估计，因此它们只适用于手势识别、手语应用等空手场景。但现实场景中手不可避免地要与周围的环境及其中的物体进行交互，因此为了进一步理解手部与物体的交互行为，本文不仅提取手部的3D姿态，同时估计交互物体的3D姿态。

与传统的仅涉及手部或仅涉及物体的3D姿态估计不同，同时估计手部与物体的姿势更具挑战性，这是因为手部与物体之间存在相互遮挡会造成关键信息丢失，严重影响估计精度。但值得注意的是，在手物交互场景中手部姿态与物体姿态存在固有的相互约束关系，有效建模这种交互依赖性能够显著提升两者的3D姿态估计精度[42-43]。现有的手物交互3D姿态估计方法主要分为两类：基于优化的方法[44-46]和基于学习的方法[4,42-43,47-50]。基于优化的方法通过考虑手和物体的接触面及内在的物理约束条件，逐步优化手和物体的姿态，但该类方法在估算手与物体之间的接触面时计算量大且耗时。基于学习的方法则采取不同的策略，设计端到端的模型从大量数据中进行学习来实现手部和物体姿态的联合估计，该方法无需进行明确的接触面计算，并能基于先验知识更高效地预测手部和交互物体的3D姿态，受到了研究人员的广泛关注并已成为主流的手部与交互物体3D姿态估计方法。2019年，Tekin等人[4]提出一种基于RGB图像的端到端手部与交互物体3D姿态估计网络。该网络通过前向传递过程同时解决手部3D姿态估计、物体3D姿态估计、物体识别及活动分类等四个任务。2020年，Huang等人[47]提出一种基于非自回归Transformer的结构化网络HOT-Net，用于联合估计手部和物体的3D姿态。该网络以精细且全面的方式对手部关节和物体角点之间的强相关性进行建模，并设计了协同姿态约束机制，以提高手部结构的物理合理性。2020年，Oberweger等人[48]设计了一个反馈回路用于纠正卷积神经网络估计错误的结果，以提升手部与物体的3D姿态估计精度，并且仅使用深度图像作为输入。2023年，Wang等人[50]提出一种密集相互注意力机制，能够模拟手和物体之间的精细依赖关系，对于每个手部节点，通过学习到的注意力机制从每个物体节点聚合特征，反之亦然，从而生成高质量且符合物理结构的3D姿态。2024年，Kuang等人[43]提出了通过施加先验条件来学习广泛的图像背景信息用于解决手与物体相互遮挡的问题，并为上下文、手部和物体解码器层提供了定制化的特征图，有助于分离这些层，降低了特征学习的复杂性。近些年的研究已经展示了多种联合理解手部与交互物体姿态的方法，这些方法分别基于深度图像、RGB图像、以及RGB-D图像，有先分别估计手部和交互物体的姿态再联合细化的方法，也有直接联合估计手部和交互物体3D姿态的方法，并通过有效利用深度特征、纹理特征、先验知识等特征信息均不同程度地提升了手部与交互物体3D姿态的估计精度。

目前，除了基于卷积神经网络[39]和Transformer神经网络[41]的手部与交互物体姿态估计方法，基于图卷积神经网络[40]的手部与交互物体姿态估计方法[42,51-53]也得到了研究者的广泛关注，主要得益于其能够以图的形式显示的建模手部与物体之间的拓扑结构和交互关系，其中，手部21个关节点和物体8个角点表示为图节点，骨骼和角点连接线表示为边。2020年，Doosti等人[51]提出了一种轻量级网络HOPE-Net，能够实时地联合估计手部和交互物体的2D与3D姿态。该网络使用两个自适应图卷积网络级联结构，其中，一个用于估计手关节和物体角点的2D坐标，另一个将获取的2D坐标转换为3D坐标。2021年，Zhuang等人[42]提出了一种基于上下文感知的图卷积网络来共同学习独立的几何先验和交互信息。该网络先从RGB图像中预测2D关键点，再经过第一个图网络分别基于手部和交互物体的拓扑结构子图估计手部和交互物体的初始3D姿态，最后将两个子图合并成一个交互拓扑图以捕捉交互信息，进一步优化3D估计结果。2024年，Hoang等人[53]提出从深度图像和RGB图像中联合学习手部和物体的纹理特征和几何特征，并基于自适应特征的深度霍夫投票策略与图卷积

网络，建模手物交互过程中的动态依赖关系，最终实现两者的联合3D姿态估计。从上述研究中总结可知，图卷积网络在手部与交互物体拓扑结构关系表示方面具有较强的优势，对于获取符合物理上合理的2D和3D姿态具有重要的作用。

综合上述研究发现，由于手部的高自由度和个体差异，以及手部与交互物体不同程度的相互遮挡等因素的影响，导致手部和交互物体姿态估计精度仍不够准确，且已有方法在2D和3D姿态估计中存在单模态数据特征表示能力受限，RGB-D数据输入特征利用不充分，以及手部与物体拓扑结构和交互依赖关系考虑较为单一等问题。因此，本文针对上述问题基于现有方法做出改进，提出一种联合多模态特征与结构感知的手物交互姿态估计方法，充分利用RGB-D图像的多模态特征、手部和交互物体自身的拓扑结构以及交互相关性等特征信息，采用两阶段的方式以获得更可靠的3D姿态。本文主要贡献包括：(1)联合深度图像和RGB图像的多模态特征，自适应的融合颜色和几何特征以缓解手部和交互物体的相互遮挡。(2)基于图结构设计引入手部、交互物体及手物交互结构感知模块对初始的2D姿态不断进行优化以获取物理结构上合理且高精度的2D姿态。(3)充分利用深度图像的空间信息和RGB图像的纹理特征，从高精度2D姿态估计结果回归得到精确的手部和交互物体3D姿态。

2　模型构建

本文网络整体框架如图1所示，采用级联式两阶段设计框架：第一阶段重点实现高精度2D姿态估计，第二阶段专注2D到3D姿态的映射预测，网络依次包含特征提取与自适应融合模块、2D姿态细化模块及3D姿态回归模块等三个部分。其中，第一部分，先通过一个堆叠的沙漏网络对深度图像和对齐深度图像的RGB图像进行特征提取，再根据手部和物体遮挡情况自适应的融合RGB-D特征获得粗2D手部姿态和物体姿态；第二部分，对获得的粗2D姿态先分别通过手部拓扑结构图卷积网络和物体拓扑结构图卷积网络进行初步的物理合理性优化，再通过手物交互拓扑相关性图卷积网络获取交互信息得到更精确可靠的手部与交互物体2D姿态。第三部分，将细化的2D姿态与深度图像中提取的深度信息z进行合并得到粗3D姿态，再利用RGB图像中提取的纹理特征对粗3D姿态进一步优化最终得到手部与交互物体的精确3D姿态。

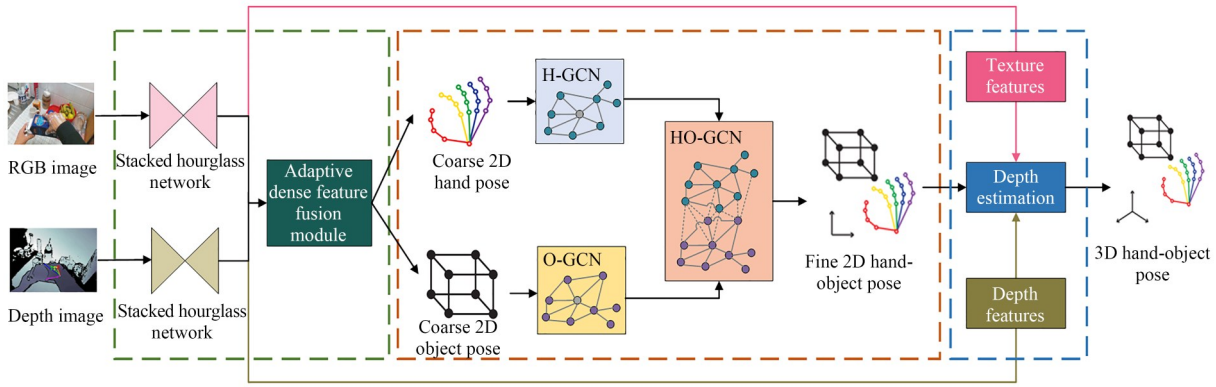


图1　网络整体框架

Fig. 1　Framework of the overall network

2.1　多模态特征提取模块

在手物交互场景中手部与物体接触导致相互遮挡，因此手部与交互物体姿态估计是一个更复杂，更具挑战性的任务。现有的主流视觉传感器（如RGB相机和深度相机）均存在视角局限性，难以通过单一观测视角完整捕捉所有手部关节点，而通过多模态数据融合的联合估计方法，可有效降低单视角条件下手物交互姿态估计的复杂度。因此，本文将单目RGB-D图像作为输入，利用改进的2个相同结构的堆叠沙漏网络[54]

分别提取深度图像的几何特征与RGB图像的纹理特征并通过融合实现信息互补，网络结构如图2所示。网络采用编解码结构，在编码阶段成组使用瓶颈残差结构(Bottleneck Residual Module，BRES)和最大池化(Max Pooling，MP)，以两倍速率逐级降低分辨率获得深层语义信息；在解码阶段成组使用BRES和上采样(Up Pooling，UP)，以两倍速率逐级恢复分辨率，其中上采样利用双线性插值法。同时，在解码阶段，还要将编码阶段相同分辨率的特征图经过BRES模块处理后，通过跳跃连接与解码阶段同分辨率的特征相加，以捕捉不同尺度特征图的手部和交互物体的特征信息。另外，在特征输出阶段，该特征经过两个分支处理融合后得到最终输出结果。其中，一个分支经过1×1的卷积得到相同的维度特征；另一个分支经过1×1的卷积输出得到中间层生成的热图，可计算损失，有助于网络提取更有效的手关节和物体角点相关特征，之后再经过1×1的卷积输出与上一分支相同维度特征再进行特征融合输出。

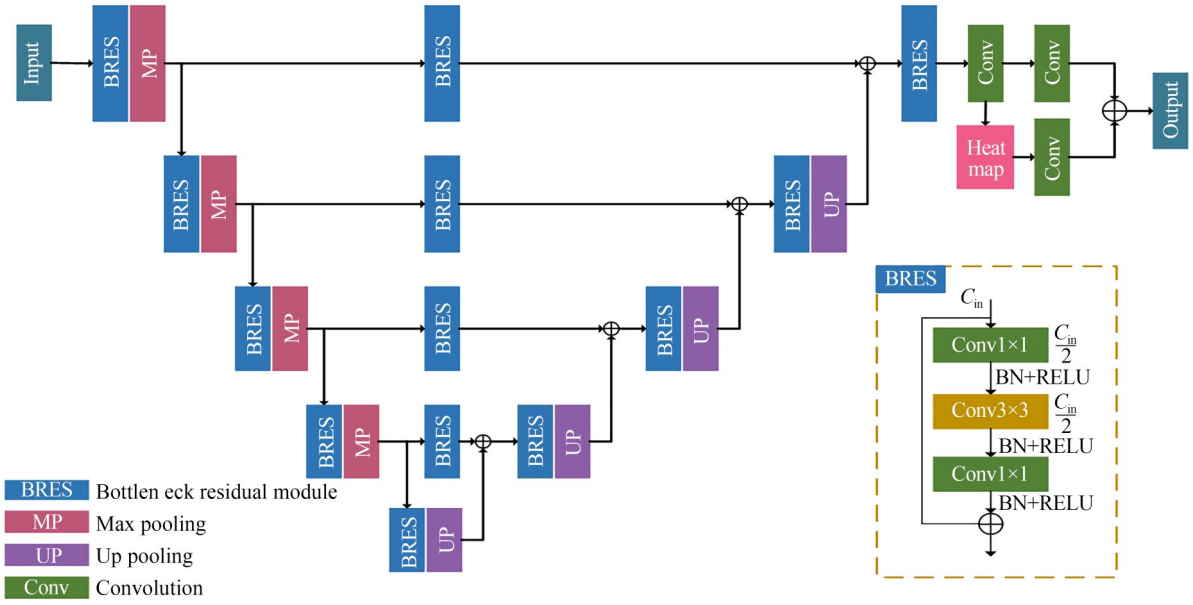


图2　特征提取模块

Fig. 2　Module of the feature extraction

2.2　自适应密集特征融合模块

深度图像和RGB图像包含与交互姿态有关的不同信息，当手部与交互物体存在不同程度的相互遮挡时，颜色特征和几何特征所提供的特征信息对精确估计2D姿态的助益是有差异的，如何合理地融合不同信息是提升姿态估计准确性的关键[55]。因此，本文在特征提取模块之后引入了自适应特征融合网络，通过学习预测每个像素特征的有用性有选择地增强和削弱两类特征信息的贡献程度，在像素级别有区分度的自适应密集融合RGB-D特征，从而实现手部与交互物体在不同遮挡情况下更精确的姿态估计结果。自适应密集特征融合模块网络结构如图3所示。

两种模态数据经过预处理输入特征提取模块后输出纹理特征 $\boldsymbol{F}_{\text{rgb}}\in\boldsymbol{R}^{H\times W\times 256}$ 和几何特征 $\boldsymbol{F}_{\text{depth}}\in\boldsymbol{R}^{H\times W\times 256}$。为了抑制背景信息的干扰，从全局图像中重点关注任务的关键区域，两种模态数据在像素级密集融合之前先经过空间注意力机制的处理，分别得到空间注意力权重矩阵 $\boldsymbol{M}_{\text{rgb}}^{H\times W\times 1}$ 和 $\boldsymbol{M}_{\text{depth}}^{H\times W\times 1}$，计算过程如公式(1)所示：

$$\begin{aligned}\boldsymbol{M}_{\text{rgb}}^{H\times W\times 1}&=\sigma\left(conv_{7\times 7}\left(avgpool\left(\boldsymbol{F}_{\text{rgb}}^{H\times W\times 256}\right),maxpool\left(\boldsymbol{F}_{\text{rgb}}^{H\times W\times 256}\right)\right)\right)\\\boldsymbol{M}_{\text{depth}}^{H\times W\times 1}&=\sigma\left(conv_{7\times 7}\left(avgpool\left(\boldsymbol{F}_{\text{depth}}^{H\times W\times 256}\right),maxpool\left(\boldsymbol{F}_{\text{depth}}^{H\times W\times 256}\right)\right)\right)\end{aligned},\tag{1}$$

其中：σ 表示激活函数，$conv_{7\times 7}$ 表示使用7×7卷积核对合并后的特征进行卷积操作，$avgpool(\ )$

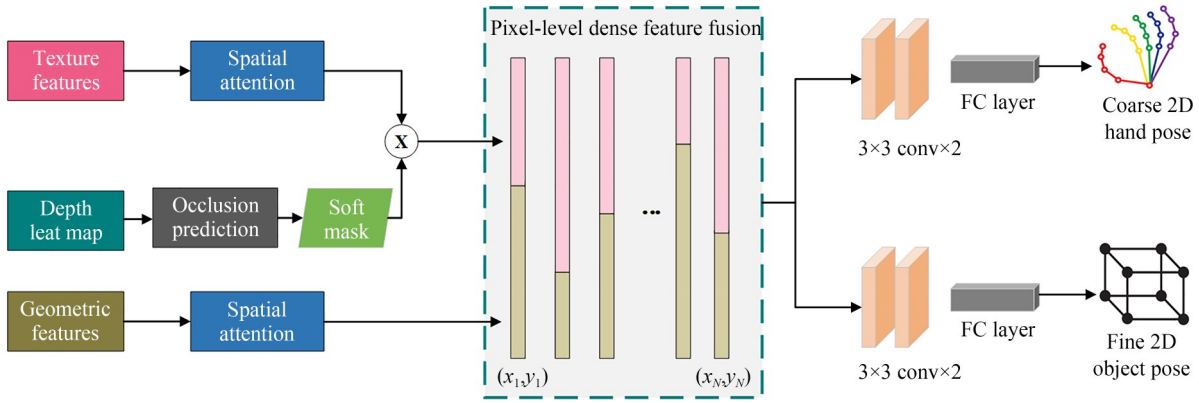


图 3 自适应密集特征融合网络

Fig. 3 Network of adaptive dense feature fusion

和 $maxpool(\ )$ 分别表示平均池化操作和最大池化操作。再将权重矩阵与输入特征融合实现自适应细化，得到最终输出特征 $\boldsymbol{G}_{\mathrm{rgb}}^{H\times W\times 256}$ 和 $\boldsymbol{G}_{\mathrm{depth}}^{H\times W\times 256}$，计算过程如公式(2)所示：

$$\left(\boldsymbol{G}_{\mathrm{rgb}}^{H\times W\times 256},\boldsymbol{G}_{\mathrm{depth}}^{H\times W\times 256}\right)=\left(\left(\boldsymbol{M}_{\mathrm{rgb}}^{H\times W\times 1}\otimes\boldsymbol{F}_{\mathrm{rgb}}^{H\times W\times 256}\right),\left(\boldsymbol{M}_{\mathrm{depth}}^{H\times W\times 1}\otimes\boldsymbol{F}_{\mathrm{depth}}^{H\times W\times 256}\right)\right). \tag{2}$$

先前的研究发现[2,52]，当手被操作物体遮挡程度达到一定阈值(可见的关节点数少于 16 个)，由于 RGB 图像的信息严重缺失会造成手部估计误差明显增大，与深度图联合估计反而会起负面作用。因此，本文在像素级密集特征融合之前对 RGB 图像进行了软掩膜处理，利用深度热图进行遮挡预测，通过计算各手部关节点深度值与真实 z 值标注的差值进行遮挡检测：当差值超过预设深度阈值(30 mm)时判定为遮挡关节。若 RGB 图像中被遮挡关节数量超过 5 个，则认为存在严重手物遮挡情况，此时软掩膜权重趋近于 0，相应抑制 RGB 特征的贡献；反之，权重趋近于 1，保留完整的 RGB 特征用于后续姿态估计。再将软掩膜与纹理特征提取分支相乘，得到新的特征图 $\hat{\boldsymbol{G}}_{\mathrm{rgb}}\in\boldsymbol{R}^{H\times W\times 256}$。接着，将经过软掩模操作的纹理特征与几何特征进行像素级密集融合后输出特征 $\boldsymbol{F}_{\mathrm{fusion}}^{H\times W\times 256}$，计算公式如式(3)所示：

$$\boldsymbol{F}_{\mathrm{fusion}}^{H\times W\times 256}=DenseFusion\left(\hat{\boldsymbol{G}}_{\mathrm{rgb}},\boldsymbol{G}_{\mathrm{depth}}\right). \tag{3}$$

最后，分别经过两个 3×3 卷积和一个全连接层支路预测输出手部和交互物体的粗 2D 姿态估计结果。

2.3 结构感知模块

精确估计手部和交互物体的 2D 姿态对于提升 3D 姿态的精度具有重要的作用，而手部在手物交互中起主导作用，正确的手部姿态估计对物体的姿态估计也有很大的影响。尽管手部骨架的树状铰链结构在二维投影与三维空间位置间存在固有几何约束，但由于个体手型差异的复杂性，导致这种关系难以显式建模，为此文中提出通过学习来对这种非线性映射关系进行拟合。在 3D 姿态回归前，构建包含手部关节、物体角点及两者交互关系的图结构进行物理特征学习，从而实现对获取的 2D 姿态进一步优化。结构感知模块首先通过两个子图网络分别捕获手和物体各自的结构，然后将两个子图合并为一个整体图，以挖掘交互信息。手部和物体的物理结构如图 4 所示，手部使用 21 个关节点表示，物体使用矩形框 8 个角点表示。手部结构特征可分为显式和隐式两个维度。其中，显式特征体现为解剖学连接关系，通过对骨骼长度、方向及夹角的建模，能够准确捕捉手部尺寸范围和空间朝向等几何属性；隐式特征则反映运动学约束关系，即不同手指的同序关节在运动功能结构方面呈现出一定的对称性(水平方向进行层级划分，可明确界定出 5 个关节层级，同一层级的关节，其运动活动范围具有一致性，易产生趋同的运动趋势)、不同手指的运动之间也存在一定的依赖关系(例如：大拇指和食指的自由度和灵巧度最高，其余三指存在相互制约关系)，通过从数据中学习内在的语义信息和相应的特征表示邻接矩阵，可进一步

优化手部2D姿态。物体的结构特征表示为长方体，可通过几何结构信息对物体2D姿态实现优化。在手物交互关系中，由于物体的形状和大小决定了人手的抓握姿势，当手指的指尖位置与手指的底部位置确定之后，手和物体的姿态信息的解空间就已经约束到了一定范围，因此通过子图合并后的整体图学习手物交互关系对提升2D姿态估计精度同样具有重要的价值。

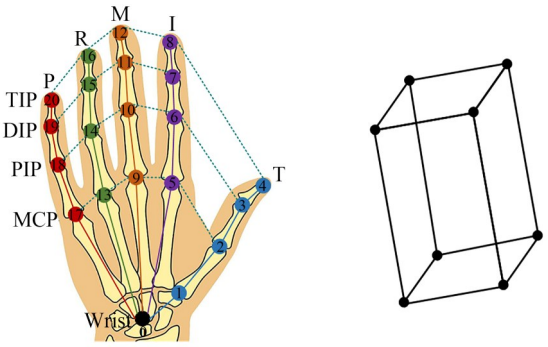


图4　手部和物体的物理结构

Fig. 4　Physical structures of hands and objects

2.4　3D姿态回归模块

在3D姿态回归过程中，考虑到RGB图像的深度模糊性，而深度图像素值直接表征场景点到相机光心的欧氏距离，与关节点/角点的z坐标存在线性映射关系，因此，网络架构仅依赖深度图来回归手部关节和物体角点的z轴向坐标。3D姿态回归网络如图5所示，获取的几何特征通过空间注意力抑制背景信息，得到增强的手部与交互物体z方向坐标，而对于xy平面坐标的估计，网络同时利用RGB特征和深度特征进行多模态融合，随后将融合后的2D坐标预测与深度图回归的z坐标整合得到粗3D姿态。由于深度图像易受分辨率和噪声的干扰，导致手部和交互物体边缘结构深度值的准确性，而纹理特征能够明确区分手部和交互物体，有助于更准确地识别两者的形状与姿态，因此，本文利用纹理特征对得到的粗3D姿态进行优化约束，通过多层感知机(Multi-Layer Perceptron，MLP)将纹理特征与粗3D姿态进行融合与处理，进一步细化和校正手部与交互物体的3D姿态。

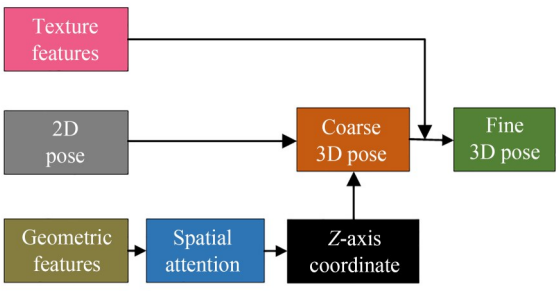


图5　3D姿态回归网络

Fig. 5　Network of 3D pose regression

2.5　损失函数

为了获得更精确的手部与交互物体3D姿态，使用多组损失函数来监督网络训练，总的损失函数$\mathcal{L}_{\text{total}}$包括热图损失$\mathcal{L}_{\text{hm}}$，2D姿态估计损失$\mathcal{L}_{\text{2D}}$及3D姿态估计损失$\mathcal{L}_{\text{3D}}$：

$$\mathcal{L}_{\text{total}}=\lambda_{\text{hm}}\mathcal{L}_{\text{hm}}+\lambda_{\text{2D}}\mathcal{L}_{\text{2D}}+\lambda_{\text{3D}}\mathcal{L}_{\text{3D}},\tag{4}$$

其中，λ_{hm}，λ_{2D}，λ_{3D}均为超参数，实验中依次设定为0.1，0.1，1。

热图损失：由于网络特征提取模块包含深度图像和RGB图像两个分支，因此，热图损失$\mathcal{L}_{\text{hm}}$包含深度热图损失$\mathcal{L}_{\text{hm-D}}$和RGB热图损失$\mathcal{L}_{\text{hm-R}}$两个部分，二维热图损失可表示为：

$$\mathcal{L}_{\text{hm}}=\mathcal{L}_{\text{hm-D}}+\mathcal{L}_{\text{hm-R}}=\sum_{i=1}^{N}\left\|\hat{D}_i-D_i\right\|_2^2+\sum_{j=1}^{M}\left\|\hat{R}_i-R_i\right\|_2^2,\tag{5}$$

其中：$\hat{D}_i$和$\hat{R}_i$分别为深度热图和RGB热图的估计值，D_i和R_i分别为深度热图和RGB热图的真值。

2D姿态估计损失：手部与交互物体二维姿态估计部分包含粗2D姿态估计和细2D姿态估计，因此2D姿态估计损失$\mathcal{L}_{\text{2D}}$也包含粗2D姿态估计损失$\mathcal{L}_{\text{2D-C}}$和细2D姿态估计损失$\mathcal{L}_{\text{2D-F}}$两个部分，可表示为：

$$\mathcal{L}_{\text{2D}}=\mathcal{L}_{\text{2D-C}}+\mathcal{L}_{\text{2D-F}}=\sum_{i=1}^{N}\left\|\hat{P}_i^{\text{2D}}-P_i^{\text{2D}}\right\|_2^2+\sum_{j=1}^{M}\left\|\hat{P}_j^{\text{2D}}-P_j^{\text{2D}}\right\|_2^2,\tag{6}$$

其中：$\hat{P}_i^{\text{2D}}$和$\hat{P}_j^{\text{2D}}$分别为二维关节点和角点的粗估计值和细估计值，P_i^{2D}和P_j^{2D}分别为二维关节点和角点的真值。

3D姿态估计损失$\mathcal{L}_{\text{3D}}$可表示为：

$$\mathcal{L}_{\text{3D}}=\sum_{i=1}^{N}\left\|\hat{P}_i^{\text{3D}}-P_i^{\text{3D}}\right\|_2^2,\tag{7}$$

其中：$\hat{P}_i^{\text{3D}}$为三维关节点和角点的估计值，P_i^{3D}为

三维关节点和角点的真值。

3 实　验

3.1 数据集

本文选择手部与交互物体三维姿态估计领域主流的数据集First-Person Hand Action (FPHA)[56]和HO-3D[57-58]开展了定量评估和定性评估实验，数据集的具体信息如表1所示。FPHA：第一个手物交互姿态估计数据集，在3种不同的实验场景采集了6位实验人员使用45种抓取动作抓取26个物体的视频(1 175个)。但包含6D物体姿态注释的帧较少(21 501帧)，仅涉及4类物体(牛奶瓶、盐、果汁盒、液体肥皂)，其中11 019帧用于训练，10 482帧用于评估，并将该子集定义为FPHA-HO[4,42,59]。每个帧的注释是一个6D向量，给出了每个物体的3D位移和旋转。HO-3D：由10名实验员对10个物体抓取时拍摄的68个视频序列组成，被划分为66 034个训练数据和11 524个测试数据。需要注意的是，在包含11 000个帧的测试集中，手仅标注了手腕坐标。

表1　手物交互姿态估计数据集

Tab. 1　Dataset of hand-object interaction pose estimation

Dataset	Year	Image	Hand label	Object label	Type
FPHA-HO	2018	21 501	Joint	Pose, Mesh	RGB-D
HO-3D	2021	77 558	Joint	Pose	RGB-D

3.2 评估指标

依据文献[4,42,51]等手部与交互物体三维姿态估计方法，本文分别使用手部姿态误差(Hand Pose Error, HPE)、物体姿态误差(Object Pose Error, OPE)、手部正确关键点百分比(Percentage of Correct Keypoints, PCK)以及物体正确姿态百分比(Percentage of Correct Pose, PCP)等评估指标对估计的手部与交互物体姿态结果进行评价。手部3D PCK曲线是指预测的关键点与真实关键点的距离小于给定阈值(单位为毫米)的比例；物体2D PCP曲线是指预测6D物体姿态在2D的投影点与真实标注误差小于一定阈值(单位为像素)的比例。此外，还提供了手部3D PCK曲线和物体PCP曲线下的面积(Area Under the Curve, AUC)。

3.3 实验环境及参数设置

实验软件环境具体参数为Ubuntu20.04，Python 3.10，Pytorch 2.4.0，CUDA 11.8，cuDNN 8，NVCC，VNC。实验硬件环境配置参数为10* Xeon Platinum 8352V的CPU，NVIDIA GeForce RTX 4090(显存24 G)的显卡、内存50 G。实验训练使用了Adam优化器对目标函数进行优化，批处理大小为8，训练从学习率为0.001开始，持续进行5 000个周期，每1 000个周期时会进行学习率的衰减，每次衰减率均为0.1。

3.4 不同方法的对比实验

为了验证RGBD图像作为输入的有效性和所提方法的优越性，选择了H+O，HOT-Net，HOPE-Net等均基于图卷积的不同方法进行了对比实验，在FPHA-HO测试集的对比实验结果如表2所示。从对比结果可以看出，本文方法明显优于其他方法，估计的手部姿态误差和物体姿态误差分别降低到9.62 mm和14.37 mm。其中，所提方法与Zhuang等人提出的方法相比，网络结构的主体均采用子图和合并子图的方式估计手部与交互物体的姿态，但Zhuang等人仅将RGB图像作为网络的输入数据，而本文采用RGBD图像作为网络的输入数据，使得手部姿态估计误差和物体姿态的估计误差分别降低了2.91 mm和4.72 mm，充分证明了深度图像和彩色图像特征的信息互补，对手部与交互物体三维姿态的精确估计确实具有重要的作用。

表2　FPHA-HO数据集的测试结果

Tab. 2　Test results of the FPHA-HO dataset (mm)

Method	HPE	OPE
H+O[4]	15.81	24.98
HOT-Net[47]	15.18	21.37
HOPE-Net[51]	15.04	21.33
Zhang[59]	14.97	23.07
Zhuang[42]	12.53	19.09
Ours	**9.62**	**14.37**

图6为FPHA-HO数据集在不同阈值下的手部3D PCK曲线。从图中可以看出在给定的0～80 mm的阈值范围内，本文方法整体明显优于其他手部与交互物体姿态估计方法，手部姿态估计

精度具有更高的准确率。其中，在0～15 mm的阈值范围内，本文方法正确3D手部姿态的百分比明显高于其他方法，在15～45 mm的阈值范围内，本文方法与Zhuang等人提出的最优方法的结果接近，但仍比其他方法具有更高的精度。因此，从上述分析可知，本文方法对3D手部姿态估计结果的精度调优具有明显的优势，能够有效约束手物交互姿态估计场景中手部物理合理性的估计。

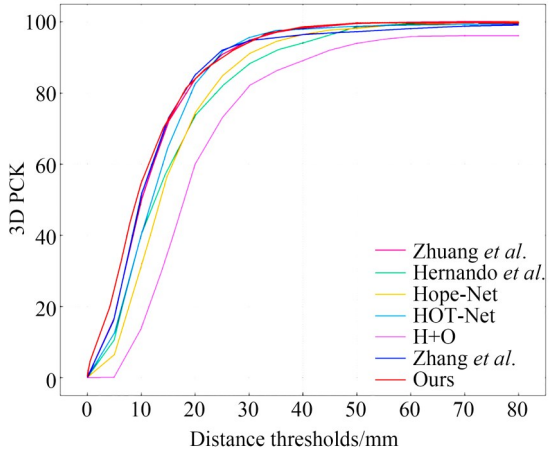


图6　FPHA-HO数据集手部3D PCK曲线

Fig. 6　3D PCK curve of hands in the FPHA-HO dataset

图7为FPHA-HO数据集在不同阈值下的物体2D PCP曲线。从图中可以看出在给定的0～50 pixels的阈值范围内，本文方法在2D投影中不同像素阈值下正确物体姿态的百分比始终高于

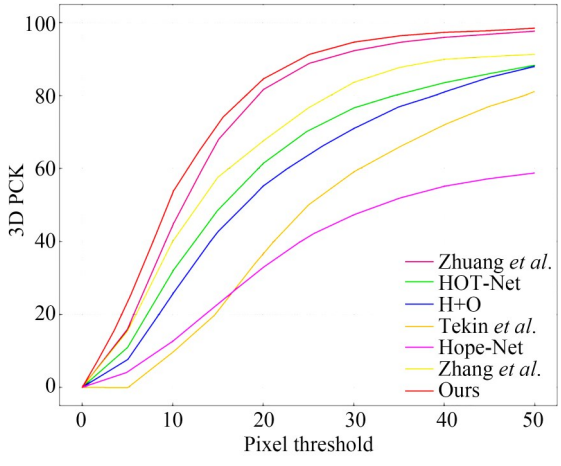


图7　FPHA-HO数据集物体2D PCP曲线

Fig. 7　2D PCP curve of objects in the FPHA-HO dataset

其他手部与交互物体姿态估计方法，在40～50 pixels阈值的范围内，与Zhuang等人提出的方法趋于接近，充分说明本文方法对物体姿态估计精度的提升具有显著的优势。

为了验证本文方法在不同手物交互操作场景的姿态估计性能，在FPHA-HO数据集选择了3位实验人员对果汁瓶、液体肥皂、牛奶瓶等不同物体的交互操作场景，开展了定量对比实验，实验结果如表3所示。从对比结果中可以看出，本文方法相比于HOPE-Net和Zhuang等人提出的方法，在三类物体交互操作场景中的手部和物体姿态估计误差均取得了最优的结果。其中，在果汁瓶的交互操作场景中，手部和交互物体的姿态估计结果相比于平均姿态估计结果仍要低于1.28 mm和2.9 mm，但在液体肥皂和牛奶瓶的交互操作场景中，手部和交互物体的姿态估计误差相比于平均估计误差均有不同程度的升高，充分说明了不同形状的物体对手的操作方式有很大的影响，从而导致遮挡程度也不尽相同。因此，本文采用RGB-D特征实现信息互补、以及利用结构感知模块实现物理合理性优化，对提升网络性能、降低姿态估计误差具有重要的作用。

表3　不同交互操作场景的测试结果

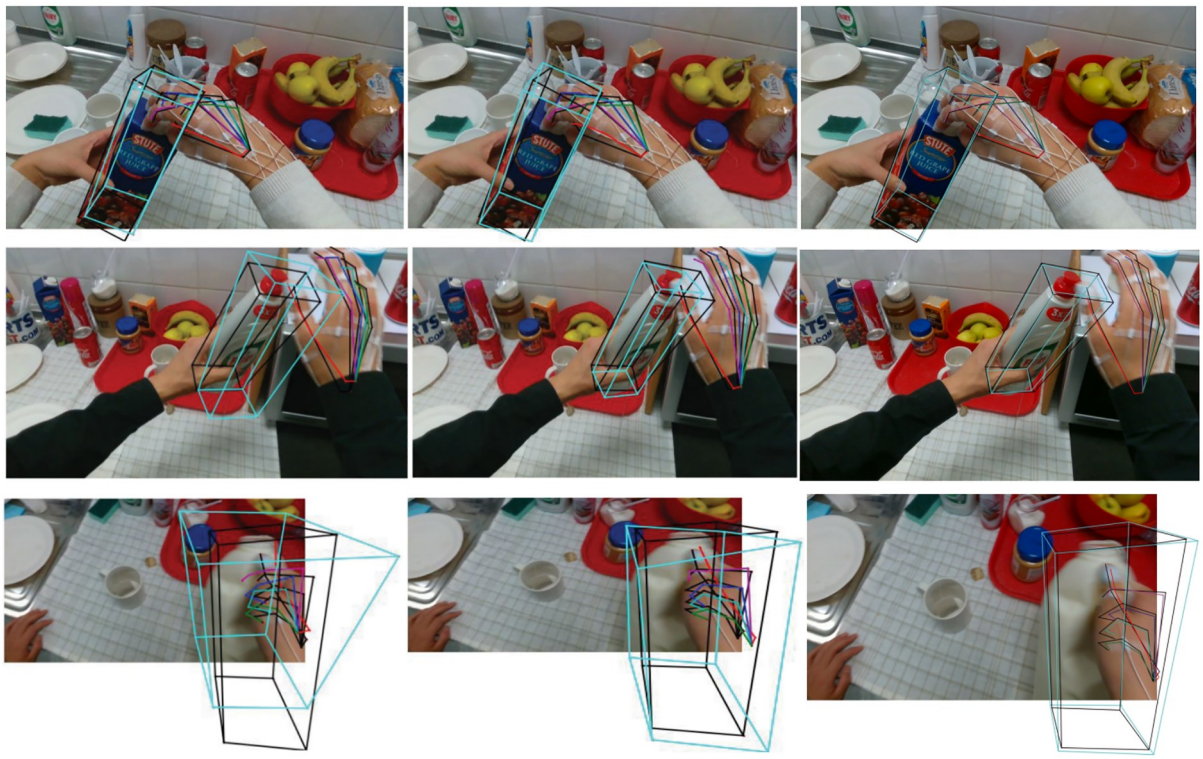
Tab. 3　Test results of the different interaction scenarios

(mm)

Method	Juice		Liquid Soap		Milk	
	HPE	OPE	HPE	OPE	HPE	OPE
HOPE-Net[51]	12.27	19.40	14.92	22.57	18.12	27.03
Zhuang[42]	9.82	17.54	12.19	18.23	14.87	19.60
Ours	**8.34**	**11.47**	**9.98**	**13.25**	**12.24**	**15.25**

为了可视化地展示不同手物交互操作场景的姿态估计效果，开展了定性对比实验，结果如图8所示，其中，黑色线条标注的为真值，彩色线条标注的为预测值(彩图见期刊电子版)。从实验结果中可以看出，相比于HOPE-Net，Zhuang等人提出的方法，本文方法能够更好地捕捉手部和交互物体的3D姿态，尤其是对于物体形状而言。在第一行“打开果汁瓶”的交互操作图片中，所有方法都能较为精准的估计手物姿态，但本文方法在手指PIP和DIP关节点及物体角点的估计结果更为准确，因此手部骨架和物体矩形框更

接近黑色真值;在第二行“打开液体肥皂”的交互操作图片中,由于手部出现了模糊现象,因此手物姿态估计结果出现了不同程度的误差,但本文方法仍然获取到了相对精确的姿态;在第三行“倒牛奶”的交互操作图片中,手和物体超出了图像边界,但本文方法对物体姿态的估计更接近真值。因此,通过上述定性实验结果分析可知,融合深度图像和彩色图像的特征,能够有效且稳定地提升手部与交互物体的姿态估计准确性;通过先分别学习手部和交互物体的先验,然后再联合学习手物交互特征,能够更稳定且合理地估计出手部和交互物体的3D姿态。此外,图9中还展示了FPHA-HO数据集中其他一些手部与物体交互操作的定性实验结果。从图中可以看出,在不同程度的相互遮挡和复杂背景下,本文方法仍然可以根据空间特征和纹理特征的信息互补,以及在结构感知模块学习到的手和物体的几何信息做出准确的预测。



(a) HOPE-Net　(b) Zhuang　(c) Ours

图8　FPHA-HO数据集定性对比实验

Fig. 8　Qualitative comparison experiment of the FPHA-HO dataset

为了验证所提方法在不同数据集的鲁棒性,在FPHA-HO和HO-3D基准数据集上进行了对比实验,结果如表4所示,将本文方法与HOT-Net及Zhang等人提出的手物交互姿态估计方法进行定量比较,评价指标包括手部姿态的3D PCK曲线下的AUC值和物体姿态在二维投影平面的PCP曲线下的AUC值。另外,由于HO-3D数据集测试集中的手只标注了手腕(没有完整的手部标注),因此,表中结果仅针对手腕关键点。从表中可以看出,本文方法在FPHA-HO数据集的3D PCK曲线和2D PCP曲线下的AUC值均达到了最高,与Zhang等人提出的方法相比,AUC值分别提升了1.2%和9.4%;而在HO-3D数据集的实验结果中,本文方法仍比其他两个方法具有显著的优势,3D PCK曲线下的AUC值和2D PCP曲线下的AUC值分别达到了83.7%和60.1%。综合实验结果和数据分析可知,本文方法在不同数据上的性能均有一定程度的提升,进一步验证了本文方法对手物交互姿态估计任务在准确性、泛化性等方面的提升是有效的。

图 9　不同手物交互场景的定性实验结果

Fig. 9　Qualitative experimental results of different hand-object interaction scenarios

表 4　不同数据集的 AUC 实验结果

Tab. 4　AUC experimental results of different datasets

Dataset	Method	AUC on 3D PCK	AUC on 2D PCP
FPHA-HO	HOT-Net[47]	0.829	0.595
	Zhang[59]	0.839	0.654
	Ours	**0.851**	**0.748**
HO-3D	HOT-Net[47]	0.819	0.567
	Zhang[59]	0.805	0.583
	Ours	**0.837**	**0.601**

3.5　消融实验

为了评估各个模块的贡献，在 FPHA-HO 数据集上设计了消融实验，定量分析了不同模块对手物姿态估计性能的影响，基线模型采用输入图像为 RGB 的 Zhuang 等人提出的方法进行对比，实验结果如表 5 所示，表格中的“w/o”表示去除该模块。对表中的实验数据分析可知，首先，输入图像仅为单模态的 RGB 图像或深度图像，相比于基线模型手部姿态估计误差分别降低了 0.3 mm 和 1.15 mm，物体姿态估计误差分别降低了 0.92 mm 和 2.57 mm，因此深度图像的空间信息对手物交互姿态估计任务更有帮助；其次，将输入的多模态特征经过等比例融合，相比于本文方法手部姿态估计误差和物体姿态估计误差分别升高了 1.29 mm 和 2.07 mm，充分说明 RGB-D 特征的自适应融合有助于深度特征和纹理特征的合理利用，从而实现更精准的手物交互姿态估计；接着，对获取的手部和交互物体粗 2D 姿态分别经过结构感知模块优化后未进行交互学习，以验证手物交互学习模块的有效性，相比于本文方法手部姿态估计误差和物体姿态估计误差分别升高了 1.45 mm 和 2.12 mm，因此通过图卷积网络学习建模手部和物体的交互操作，对提升手物交互姿态估计结果具有重要的意义；最后，采用获取到的细 2D 手部与交互物体姿态直接回归得到 3D 姿态代替本文提出的 3D 姿态优化模块，相比于本文方法手部姿态估计误差和物体姿态估计误差分别升高了 0.23 mm 和 0.91 mm，表明了采用深度信息和纹理特征分别对手部和交互物体 3D 姿态的获取和优化是非常有效的，能够显著提升手物交互姿态估计任务的准确性。

表 5　网络不同模块的实验结果

Tab. 5　Experimental results of different modules of the network　(mm)

Method	HPE	OPE
Zhuang[42]	12.53	19.09
w/o Depth	12.23	18.17
w/o RGB	11.38	16.52
w/o Adaptive fusion module	10.91	16.44
w/o Interactive learning module	11.07	16.49
w/o 3D Optimization module	9.85	15.28
Ours	**9.62**	**14.37**

为了评估RGB-D图像在2D(xy)和z方向的贡献度,在FPHA-HO数据集上设计了消融实验,定量分析了不同输入图像对手物姿态估计性能的影响,实验结果如表6所示。从表中可以看出,输入RGB-D图像的手部和物体姿态估计误差相比于仅输入RGB图像的手部和物体姿态估计误差,在2D(xy)方向分别降低了1.86 mm和2.69 mm,在z方向分别降低了3.5 mm和5.18 mm;而仅输入深度图像的手部和物体姿态估计误差相比于仅输入RGB图像的手部和物体姿态估计误差,在2D(xy)方向分别降低了0.41 mm和0.92 mm,在z方向分别降低了3.44 mm和4.87 mm。因此,深度图像虽在手部和物体姿态估计的z方向精度提升较为明显,但对2D(xy)方向的精度提升也有助益;而本文提出的联合多模态特征相比于单模态特征输入,对提升2D(xy)和z轴向的估计精度均有显著的作用。

表6　网络不同输入图像的实验结果

Tab. 6　Experimental results of different input images of the network　(mm)

Image	HPE			OPE		
	2D(xy)	z	3D	2D(xy)	z	3D
RGB	9.54	8.24	12.23	14.17	12.26	18.17
Depth	9.13	4.80	11.38	13.25	7.39	16.52
RGB-D	**7.68**	**4.74**	**9.62**	**11.48**	**7.08**	**14.37**

4　结　论

本文提出一种联合多模态特征与结构感知的手部与交互物体3D姿态估计方法,采用两阶段网络结构,依次包含特征提取与自适应融合、2D姿态细化、3D姿态回归三个模块。该方法利用彩色与深度图像特征实现信息互补,有效解决遮挡与复杂背景问题;基于图结构设计手部、物体及手物交互结构感知模块,辅助估计合理准确的2D姿态;将2D姿态与深度信息合并,利用纹理特征优化得到最终3D姿态。实验在FPHA-HO,HO-3D等数据集开展,结果表明,所提方法优于现有方法,手部和交互物体姿态误差分别降至9.62 mm和14.37 mm。然而,本研究仍存在一些不足之处。在数据集方面,尽管选用了FPHA-HO,HO-3D等具有一定代表性的数据集,但这些数据集的场景和物体类型仍相对有限。实际应用中,手与物体的交互场景更为复杂多样,可能涉及各种不同材质、形状和大小的物体,以及各种光照条件和环境背景。因此,当前方法在这些更复杂场景下的泛化能力还有待进一步验证和提升。

作者贡献声明:

王文润:姿态估计方法的提出和实验设计,论文构思和撰写;

党建武:论文修改和资源支持;

王阳萍:论文构思和数据分析;

任鹏百:实验验证和可视化呈现;

潘瑞:论文审核与编辑写作。

参考文献:

[1] 车云龙, 齐越. 基于深度图像的手部姿态估计综述[J]. 计算机辅助设计与图形学学报, 2021, 33(11): 1635-1648.
CHE Y L, QI Y. A survey on depth based hand pose estimation[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2021, 33(11): 1635-1648. (in Chinese)

[2] 李少东, 罗凯, 黄远智, 等. 复杂交互场景下融合关节遮挡信息的手部姿态估计研究[J]. 计算机学报, 2025, 48(5): 1212-1231.
LI S D, LUO K, HUANG Y Z, *et al*. Hand pose estimation *via* fusing joint occlusion informationin complex interaction scenarios[J]. *Chinese Journal of Computers*, 2025, 48(5): 1212-1231. (in Chinese)

[3] REN P F, CHEN Y C, HAO J C, *et al*. Two heads are better than one: image-point cloud network for depth-based 3D hand pose estimation[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, 37(2): 2163-2171.

[4] TEKIN B, BOGO F, POLLEFEYS M. H+O: unified egocentric recognition of 3D hand-object poses and interactions[C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 15-20, 2019. *Long Beach*, *CA*, *USA*. IEEE, 2019: 4511-4520.

[5] WANG T C, YANG T, DANELLJAN M, *et al*. Learning human-object interaction detection using interaction points[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 13-19, 2020. *Seattle*, *WA*, *USA*. IEEE, 2020: 4116-4125.

[6] HAN S, LIU B, CABEZAS R, *et al*. Megatrack: monochrome egocentric articulated hand-tracking for virtual reality[J]. *ACM Transactions on Graphics* (*ToG*), 2020, 39(4): 87: 1-87: 13.

[7] WANG J Y, MUELLER F, BERNARD F, *et al*. RGB2Hands[J]. *ACM Transactions on Graphics*, 2020, 39(6): 1-16.

[8] YE Z J, JIA J, XING J L. Semantics2Hands: transferring hand motion semantics between avatars [C]. *Proceedings of the* 31*st ACM International Conference on Multimedia. Ottawa ON Canada.* ACM, 2023.

[9] LIU X, YI L. Geneoh diffusion: towards generalizable hand-object interaction denoising via denoising diffusion [C]. 12*th International Conference on Learning Representations* (*ICLR*). *May* 7 - 11, 2024. *Hybrid*, *Vienna*, *Austria*. IEEE, 2024.

[10] ROGEZ G, SUPANCIC J S, RAMANAN D. Understanding everyday hands in action from RGB-D images[C]. 2015 *IEEE International Conference on Computer Vision* (*ICCV*). *December* 7-13, 2015, *Santiago*, *Chile*. IEEE, 2015: 3889-3897.

[11] XU B S, ZHENG S P, JIN Q. POV: Prompt-oriented view-agnostic learning for egocentric hand-object interaction in the multi-view world[C]. *Proceedings of the* 31*st ACM International Conference on Multimedia. Ottawa ON Canada.* ACM, 2023: 2807-2816.

[12] ZHU T Q, WU R N, LIN X B, *et al*. Toward Human-like grasp: dexterous grasping *Via* semantic representation of object-hand[C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 10-17, 2021, *Montreal*, *QC*, *Canada*. IEEE, 2021: 15721-15731.

[13] HANDA A, VAN WYK K, YANG W, *et al*. DexPilot: vision-based teleoperation of dexterous robotic hand-arm system[C]. 2020 *IEEE International Conference on Robotics and Automation* (*ICRA*). *May* 31-*August* 31, 2020. *Paris*, *France*. IEEE, 2020: 9164-9170.

[14] WANG R, KTISTAKIS S, ZHANG S W, *et al*. *POV-surgery: a dataset for egocentric hand and Tool Pose Estimation During Surgical Activities* [M]. Medical Image Computing and Computer Assisted Intervention-MICCAI 2023. Cham: Springer Nature Switzerland, 2023: 440-450.

[15] LI C, ZHANG R, WONG J, *et al*. Behavior-1k: A benchmark for embodied Ai with 1,000 everyday activities and realistic simulation[C]. *Conference on Robot Learning*. PMLR, 2023: 80-93.

[16] TAN M K, ZHUANG Z W, CHEN S T, *et al*. EPMF: efficient perception-aware multi-sensor fusion for 3D semantic segmentation [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(12): 8258-8273.

[17] XIAO Z, WANG T, WANG J, *et al*. Unified human-scene interaction via prompted chain-of-contacts[C]. 12*th International Conference on Learning Representations* (*ICLR*). *May* 7 - 11, 2024. *Hybrid*, *Vienna*, *Austria*. ICLR, 2024.

[18] GOUDIE D, GALATA A. 3D Hand-object pose estimation from depth with convolutional neural networks[C]. 2017 12*th IEEE International Conference on Automatic Face & Gesture Recognition* (*FG* 2017). *May* 30 - *June* 3, 2017, *Washington*, *DC*, *USA*. IEEE, 2017: 406-413.

[19] WAN C D, PROBST T, GOOL L V, *et al*. Dense 3D regression for hand pose estimation[C]. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition. June* 18-23, 2018, *Salt Lake City*, *UT*, *USA*. IEEE, 2018: 5147-5156.

[20] YUAN S X, GARCIA-HERNANDO G, STENGER B, *et al*. Depth-based 3D hand pose estimation: from current achievements to future Goals[C]. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. June 18-23,

2018, *Salt Lake City*, *UT*, *USA*. IEEE, 2018: 2636-2645.

[21] FANG L P, LIU X Y, 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[M]. *Computer Vision – ECCV* 2020. *Cham*: *Springer International Publishing*, 2020: 120-137.

[22] 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.

[23] ZIMMERMANN C, BROX T. Learning to estimate 3D hand pose from single RGB images[C]. 2017 *IEEE International Conference on Computer Vision* (*ICCV*). *October* 22-29, 2017, *Venice*, *Italy*. IEEE, 2017: 4913-4921.

[24] IQBAL U, MOLCHANOV P, BREUEL T, *et al*. *Hand Pose Estimation Via Latent* 2.5*D Heatmap Regression* [M]. Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 125-143.

[25] PANTELERIS P, OIKONOMIDIS I, ARGYROS A. Using a single RGB frame for real time 3D hand pose estimation in the wild[C]. 2018 *IEEE Winter Conference on Applications of Computer Vision* (*WACV*). *March* 12-15, 2018, *Lake Tahoe*, *NV*, *USA*. IEEE, 2018: 436-445.

[26] CAI Y J, GE L H, CAI J F, *et al*. *Weakly-supervised* 3*D Hand Pose Estimation from Monocular RGB Images*[M]. Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 678-694.

[27] BAEK S, KIM K I, KIM T K. Pushing the envelope for RGB-based dense 3d hand pose estimation *Via* neural rendering[C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 15-20, 2019, *Long Beach*, *CA*, *USA*. IEEE, 2019: 1067-1076.

[28] LIU Y, JIANG J, SUN J H. Hand pose estimation from RGB images based on deep learning: a survey[C]. 2021 *IEEE* 7*th International Conference on Virtual Reality* (*ICVR*). *May* 20-22, 2021, *Foshan*, *China*. IEEE, 2021: 82-89.

[29] LEE S, PARK H, KIM D U, *et al*. Image-free domain generalization *Via* CLIP for 3D hand pose estimation[C]. 2023 *IEEE/CVF Winter Conference on Applications of Computer Vision* (*WACV*). *January* 2-7, 2023, *Waikoloa*, *HI*, *USA*. IEEE, 2023: 2933-2943.

[30] 肖一, 刘越. 基于RGB图像的三维人手姿态估计技术综述[J]. 计算机辅助设计与图形学学报, 2024, 36(2): 161-172.
XIAO Y, LIU Y. Review on 3D hand pose estimation based on a RGB image[J]. *Journal of Computer—Aided Design & Computer Graphics*, 2024, 36(2): 161-172. (in Chinese)

[31] 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.
MA S L, LI J H, KONG D H, *et al*. 3D hand pose estimation based on double branches with multi-scale attention[J]. *Chinese Journal of Computers*, 2023, 46(7): 1383-1395. (in Chinese)

[32] JIANG X, MA X H. *Dynamic Graph CNN with Attention Module for* 3*D Hand Pose Estimation* [M]. Advances in Neural Networks-ISNN 2019. Cham: Springer International Publishing, 2019: 87-96.

[33] CHENG W C, PARK J H, KO J H. HandFoldingNet: a 3D hand pose estimation network using multiscale-feature guided folding of a 2D hand skeleton[C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 10-17, 2021, *Montreal*, *QC*, *Canada*. IEEE, 2021: 11240-11249.

[34] GAO D H, ZHANG X D, CHEN X Y, *et al*. CycleHand: increasing 3D pose estimation ability on in-the-wild Monocular Image Through Cyclic Flow [C]. *Proceedings of the* 30*th ACM International Conference on Multimedia*. *Lisboa Portugal*. ACM, 2022: 2452-2463.

[35] ZHANG X, LI Q, MO H, *et al*. End-to-end Hand mesh recovery from a monocular RGB image [C]. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 27-*November* 2, 2019. *Seoul*, *Korea* (*South*). IEEE, 2019: 2354-2364.

[36] WAN C D, PROBST T, VAN GOOL L, *et al*. Self-supervised 3D hand pose estimation through training by fitting[C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 15-20, 2019, *Long Beach*, *CA*, *USA*. IEEE, 2019: 10845-10854.

[37] SPURR A, IQBAL U, MOLCHANOV P, *et al*. *Weakly Supervised* 3D *Hand Pose Estimation Via Biomechanical Constraints*[M]. Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 211-228.

[38] 孙迪钢, 张平. 基于先验知识和网格监督的手部姿态估计[J]. 华南理工大学学报(自然科学版), 2024, 52(6): 138-147.
SUN D G, ZHANG P. Hand pose estimation based on prior knowledge and mesh supervision [J]. *Journal of South China University of Technology* (*Natural Science Edition*), 2024, 52(6): 138-147. (in Chinese)

[39] GU J X, WANG Z H, KUEN J, *et al*. Recent advances in convolutional neural networks[J]. *Pattern Recognition*, 2018, 77: 354-377.

[40] WU Z H, PAN S R, CHEN F W, *et al*. A comprehensive survey on graph neural networks[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2021, 32(1): 4-24.

[41] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[J]. *Advances in neural information processing systems*, 2017, 30.

[42] ZHUANG N, MU Y D. Joint hand-object pose estimation with differentiably-learned physical contact point analysis[C]. *Proceedings of the* 2021 *International Conference on Multimedia Retrieval. Taipei*, *Taiwan*. ACM, 2021: 420-428.

[43] KUANG Z S, DING C X, YAO H. Learning context with priors for 3D interacting hand-object pose estimation[C]. *Proceedings of the* 32*nd ACM International Conference on Multimedia. Melbourne VIC Australia*. ACM, 2024: 768-777.

[44] CAO Z, RADOSAVOVIC I, KANAZAWA A, *et al*. Reconstructing hand-object interactions in the wild[C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 10-17, 2021, *Montreal*, *QC*, *Canada*. IEEE, 2021: 12397-12406.

[45] HASSON Y, VAROL G, SCHMID C, *et al*. Towards Unconstrained Joint hand-object Reconstruction from RGB Videos[C]. 2021 *International Conference on* 3*D Vision* (3*DV*). *December* 1-3, 2021. *London*, *United Kingdom*. IEEE, 2021: 659-668.

[46] YANG L X, ZHAN X Y, LI K L, *et al*. CPF: learning a contact potential field to model the hand-object interaction[C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 10-17, 2021, *Montreal*, *QC*, *Canada*. IEEE, 2021: 11077-11086.

[47] HUANG L, TAN J C, MENG J J, *et al*. HOT-net: non-autoregressive transformer for 3D hand-object pose estimation[C]. *Proceedings of the* 28*th ACM International Conference on Multimedia. Seattle WA USA*. ACM, 2020: 3136-3145.

[48] OBERWEGER M, WOHLHART P, LEPETIT V. Generalized feedback loop for joint hand-object pose estimation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, 42(8): 1898-1912.

[49] LIU S W, JIANG H W, XU J R, *et al*. Semi-supervised 3D hand-object poses estimation with interactions in time[C]. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 20-25, 2021. *Nashville*, *TN*, *USA*. IEEE, 2021: 14687-14697.

[50] WANG R, MAO W, LI H D. Interacting hand-object pose estimation *Via* dense mutual attention [C]. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). January 2-7, 2023, Waikoloa, HI, USA. IEEE, 2023: 5724-5734.

[51] DOOSTI B, NAHA S, MIRBAGHERI M, *et al*. HOPE-net: a graph-based model for hand-object pose estimation[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 13-19, 2020. *Seattle*, *WA*, *USA*. IEEE, 2020: 6608-6617.

[52] TSE T H E, KIM K I, LEONARDIS A, *et al*. Collaborative learning for hand and object reconstruction with attention-guided graph convolution [C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 1654-1664.

[53] HOANG D C, TAN P X, NGUYEN A N, *et al*. Multi-modal hand-object pose estimation with adaptive fusion and interaction learning[J]. *IEEE Access*, 2024, 12: 54339-54351.

[54] NEWELL A, YANG K Y, DENG J. *Stacked Hourglass Networks for Human Pose Estimation* [M]. Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 483-499.

[55] WANG C, XU D F, ZHU Y K, *et al*. DenseFusion: 6D object pose estimation by iterative dense fusion[C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 15-20, 2019, *Long Beach*, *CA*, *USA*. IEEE, 2019: 3338-3347.

[56] GARCIA-HERNANDO G, YUAN S X, BAEK S, *et al*. First-person hand action benchmark with RGB-D videos and 3D hand pose annotations[C]. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. *June* 18-23, 2018, *Salt Lake City*, *UT*, *USA*. IEEE, 2018: 409-419.

[57] HAMPALI S, RAD M, OBERWEGER M, *et al*. HOnnotate: a method for 3D annotation of hand and object poses[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 13-19, 2020, *Seattle*, *WA*, *USA*. IEEE, 2020: 3193-3203.

[58] HAMPALI S, SARKAR S D, LEPETIT V. Ho-3d_v3: Improving the accuracy of hand-object annotations of the ho-3d dataset[J]. *arXiv preprint arXiv*:2107.00887, 2021.

[59] ZHANG M M, LI A, LIU H L, *et al*. Coarse-to-fine hand-object pose estimation with interaction-aware graph convolutional network[J]. *Sensors*, 2021, 21(23): 8092.

作者简介：

王文润(1990—),女,甘肃白银人,博士研究生,工程师,2013年于重庆师范大学获得学士学位,2016年于兰州交通大学获得硕士学位,主要从事人机交互及虚拟现实方面的研究。E-mail: wangwenrun@mail.lzjtu.cn

通讯作者：

党建武(1963—),男,陕西富平人,博士,教授,博士生导师,1986年于兰州铁道学院获得学士学位,1992年、1996年于西南交通大学分别获得硕士和博士学位,主要从事交通信息工程及控制、智能信息处理、图像处理等方面的研究。E-mail: dangjw@mail.lzjtu.cn