复杂交互场景下融合关节遮挡信息的手部姿态估计研究

李少东 罗 凯 黄远智 刘 熹 双 丰 高 放

(广西大学电气工程学院广西电力装备智能控制与运维重点实验室 南宁 530004)

摘 要 基于视觉的手部姿态估计是计算机视觉领域的研究热点,也是理解人类交互意图的重要途径。手部在交互中不可避免会被自身或物体遮挡造成关键信息丢失,不仅影响遮挡域手部姿态估计精度,也会降低可见域手部姿态估计精度。由于现有公开数据集缺乏大规模遮挡标签,导致专注于解决遮挡问题的研究成果不足。受此启发,本文旨在通过分类手部遮挡关节,提高遮挡下手部姿态估计精度。面向复杂交互场景,以手部MANO模型和拓扑结构为基础,将手部分割为独立运动单元集合,并基于相机成像原理首次提出单手、双手和手物交互下关节遮挡判别器,能够为公开数据集制作遮挡标签。为了表明抑制遮挡关节对提高手部姿态估计精度的重要性,本文融合关节遮挡与可见性提出动态特征选择模块,并且级联于重要研究成果POEM网络上,提出了融合关节遮挡信息的手部姿态估计网络。此外,本文基于增强现实建立了具有交互意图的手部遮挡数据集HODA,在虚拟物体引导下完成预抓取和抓取运动,既能实时反馈真实操作状态,又有效地避免物体遮挡影响。为了丰富数据集,本文采用文本引导的扩散模型为手部图像生成自然且连贯的背景。在实验环节,通过7个公开数据集和HODA验证了关节遮挡分类方法的准确率超过95.07%;基于遮挡关节数量统计将3个数据集划分为无遮挡、轻微遮挡和重度遮挡,以此验证遮挡数量对手部姿态估计的不利影响;利用对比实验和权重矩阵消融实验验证了融合关节遮挡信息的手部姿态估计网络在DexYCB、HanCo和HO3D数据集上达到了先进水平;基于泛化性、相似性实验以及扩充数据集实验验证了HODA数据集的有效性。

关键词 关节遮挡分类;手部姿态估计;自遮挡;双手遮挡;物体遮挡;手部数据集;扩散模型

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2025.01212

Hand Pose Estimation via Fusing Joint Occlusion Informationin Complex Interaction Scenarios

LI Shao-Dong LUO Kai HUANG Yuan-Zhi LIU Xi SHUANG Feng GAO Fang

(Guangxi Key Laboratory of Intelligent Control and Maintenance of Power Equipment, School of Electrical Engineering, Guangxi University, Nanning 530004)

Abstract Vision-based hand pose estimation has become a crucial research topic in the field of computer vision and plays an essential role in understanding human interaction intent. During interactions, the hand is inevitably occluded by either itself or external objects, resulting in the loss of image key information. This occlusion degrades the accuracy of hand pose estimation not only in the occluded regions but also in the visible regions. Moreover, the lack of large-scale occlusion annotations in existing publicly available datasets results in insufficient research focused on addressing the occlusion problem. Inspired by this, we propose a joint occlusion classification meth-

收稿日期:2024-08-13;在线发布日期:2025-03-04。本课题得到广西自然科学基金青年基金项目(No.2023GXNSFBA026069)、广西自然科学基金面上项目(No.2025GXNSFAA069931)、广西研究生教育创新项目(No.YCSW2023056)资助。**李少东**,博士,助理教授,主要研究领域为手部姿态估计、机器人长时序操作。E-mail:lishaodong@gxu.edu.cn。**罗 凯**,硕士研究生,主要研究领域为手部姿态估计。**黄远智**,本科生,主要研究领域为机器人。**刘 熹(通信作者)**,博士研究生,主要研究领域为手部姿态估计。E-mail:2112401019@st.gxu.edu.cn。**双 丰**,博士,教授,主要研究领域为机器视觉。**高 放**,博士,教授,主要研究领域为计算机视觉。

od to improve the accuracy of hand pose estimation under occlusion. For complex interaction scenarios, based on the MANO hand model and topology, the hand is divided into independent motion sections that have rigid body motion properties. Then discriminators based on the bone-eye angle strategy and the projection strategy are first presented to classify one-hand self-occlusion, two-hand mutual-occlusion and object occlusion by principle of camera imaging. Discriminators can be used to create occlusion labels for most public datasets, significantly enriching the dataset's annotations. To verify the effect of suppressing occluded joints for improving hand pose estimation accuracy, we integrate joint occlusion and joint visibility into a dynamic feature selection module based on POEM and present a hand pose estimation network fusing joint occlusion information. The network is implemented by cascading feature extraction, dynamic feature selection, and feature utilization module. It can select only strongly related features for utilization and filter out weakly related features, which reduces the influence of occluded joints on the accuracy of hand pose estimation, thereby improving the accuracy. In addition, based on the occlusion classification method and augmented reality technology, we propose a hand occlusion dataset with occlusion labels, named HODA, where the joint occlusion classification method automatically annotates and augmented reality is used to simulate grasping interactions. Under the guidance of virtual objects, the data collection process is divided into pre-grasp and grasp motions, which not only provides real-time feedback on actual operations but also effectively avoids the influence of object occlusion during the grasping motions. A natural and coherent background is crucial for the dataset. To address the problems of unnatural position and semantic incoherence, we use a text-guided diffusion model to generate a natural and coherent background for hand images based on the content, style, and semantics of existing images. In the experiment, the proposed joint occlusion classification method achieves an impressive accuracy of over 95.07%, as verified by seven public datasets and HODA covering one-hand, two-hand, and hand-object interactions; three datasets are categorized into non-occlusion, slight occlusion, and severe occlusion based on the number of occluded joints, to validate the adverse effects of increasing the number of occluded joints; through comparative experiments and weight matrix ablation experiments, the hand pose estimation network fusing joint occlusion information outperforms existing methods, realizes satisfactory performance on the HO3D dataset. Especially on DexYCB and HanCo datasets, the network achieves the state-of-the-art performance; the effectiveness of the HODA is confirmed by the generalizability, similarity, and complementary dataset experiments.

Keywords　joint occlusion classification; hand pose estimation; self-occlusion; mutual-occlusion; object occlusion; hand dataset; diffusion model

1　引　言

基于视觉的手部姿态自然、直观地揭示了人类交互意图[1-3]。关于手部姿态估计的研究已经取得显著进展[4],精度可达 4.76 毫米[5]。然而遮挡会造成手部关键信息丢失,严重影响手部姿态估计精度,物体遮挡时误差通常超过 10 毫米[6]。因此,遮挡下如何提高手部姿态估计精度是亟待解决的难题。

手部交互时不可避免存在遮挡现象。根据交互场景可以将遮挡典型划分为单手自遮挡[7]、双手互遮挡[8-9]和物体遮挡[10]。单手自遮挡是指手部关节被手自身遮挡,主要发生在简单的自由交互中,如增强现实(Augmented Reality, AR)操作[11]。双手互遮挡是指一只手的关节被另一只手遮挡,这种情况经常出现在复杂的自由交互中,如手势交流[12]。物体遮挡是指手被交互的物体遮挡,这种情况不可避免地发生在手－物交互中,如物体抓取[13]。事实上,遮挡会造成手部关键信息丢失,不仅严重降低遮挡关节估计精度,也会影响可见关节估计精度,尽管

这种影响会随着与被遮挡关节距离增加而减弱，可仍然会影响手部全局估计精度[14-15]。这可以简单地用深度学习机理来解释[16-17]，输入状态中包含的被遮挡关节信息会通过编码和解码完成高维—低维和低维—高维传递，从而影响手部全局信息。因此，增强手部信息成为一种提高估计精度的普遍思路。Cheng 等人[5]利用多视角策略增强手部信息，提高手部姿态估计精度，但是多视角信息冗余、计算复杂[18-21]。Madadi 等人[22]和 Chen 等人[23]采用手部历史运动轨迹增强当前帧手部信息，对于手部姿态精准估计也起到了一定效果，但是时序信息往往也会造成误差传递，在遮挡严重时能力有限。

此外，也有少量研究专注于提高遮挡下手部姿态估计精度。JoonKyu 等人[24]将图像中可见区域作为主要特征，而遮挡区域作为次要特征，通过把主要特征注入给次要特征来增强被遮挡信息，进而提高手部姿态估计精度，但是随着次要特征占比增加，精度提升效果显著下降。Xu 等人[25]以手指为基本单元划分遮挡区域，指导帧间手指信息融合，提高手部姿态估计精度，但是将手指作为基本单元并不能真实反映局部遮挡信息，故提升效果有限。

影响遮挡下手部姿态估计精度提高的另一个重要因素是缺乏大规模针对遮挡类的手部数据集。人手自由度高、手指相似性强等原因造成人工注释代价极大，尤其是存在遮挡时，难以批量化手动标记。据作者团队所知，还没有专注于遮挡类的大规模手部公开数据集，这也严重制约了遮挡下手部姿态估计研究[26]。

受到上述启发，分类手部遮挡和可见关节，抑制遮挡关节对可见关节的影响，有助于解决遮挡问题；构建大规模专注于遮挡的手部数据集，从“源头”上提供丰富的训练资源，亦有助于解决遮挡问题。本文面向复杂交互场景提出关节遮挡分类方法（下文简称为分类方法），如图 1 所示。该方法利用手部结构先验知识和手部模型分割，根据相机成像原理采用三种判别器分别对单手自遮挡、双手互遮挡和物体遮挡下的关节分类。本文针对三种场景选择了 7 个具有代表性公开数据集，再结合本文提出的数据

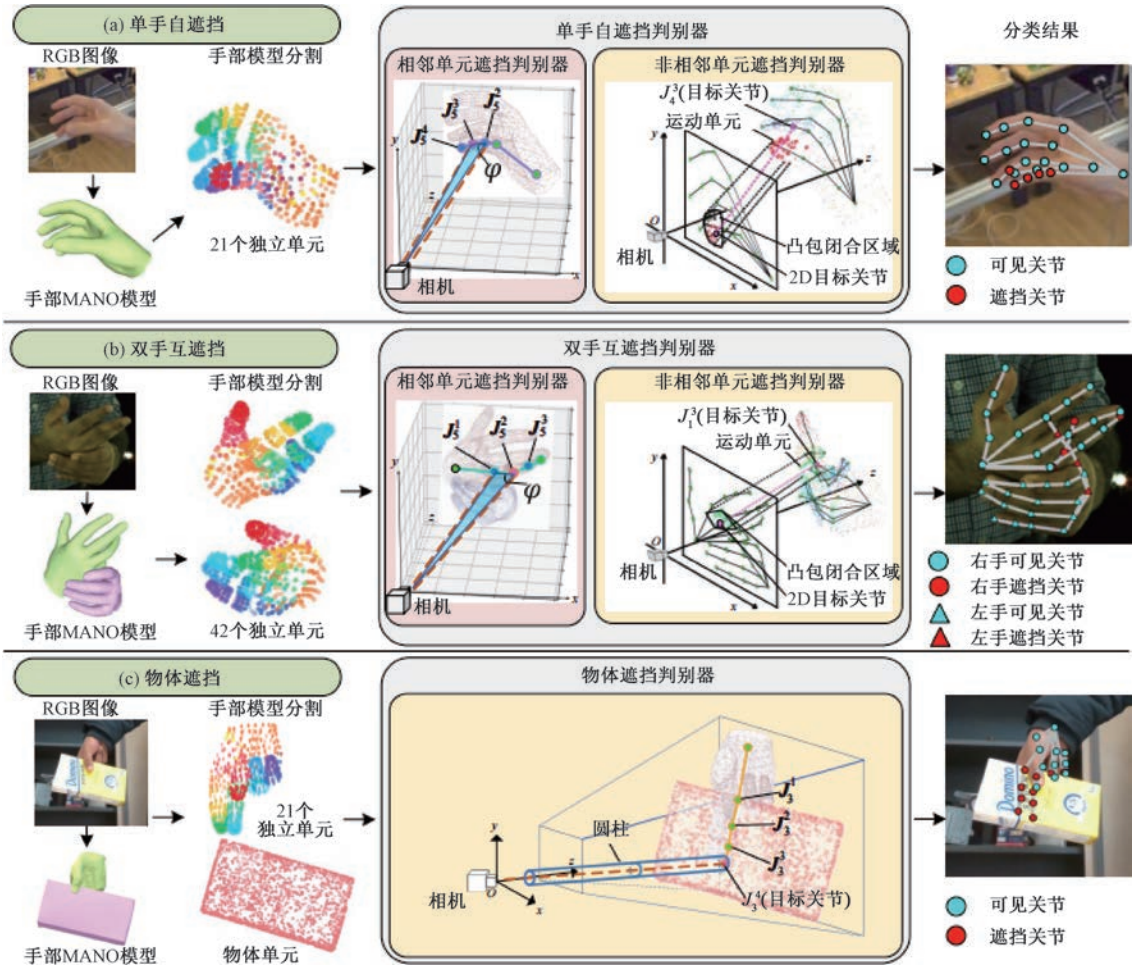


图 1 关节遮挡分类方法框架

集,开展遮挡分类准确性验证实验,准确率均超过95.07%。然后根据遮挡关节数量将3个公开数据集划分成无遮挡、轻微遮挡和重度遮挡,定量评估遮挡关节数量对精度的不利影响。为了进一步提高遮挡下的估计精度,本文设计了动态特征选择模块,尝试在重要研究成果上融合遮挡与可见关节信息。本文选择以文献[18]的POEM网络为基础,提出了融合关节遮挡信息的手部姿态估计网络。POEM通过使用多视角信息交互的方式获取3D关节和网格点位置,在常用的公开数据集DexYCB、HO3D上均取得了优于之前方法的表现。但是POEM认为每个视角信息对于获取3D关节位置具有相同的作用,使用全部信息进行估计不仅存在大量冗余,还未考虑到遮挡问题对估计精度的影响。因此,本文通过关节遮挡状态构建遮挡权重矩阵进行动态特征选择,仅选择可见关节特征进行利用。既避免了信息冗余,又避免了遮挡关节特征干扰。在单手DexYCB、HO3D和HanCo数据集上进行测试,精度相比于目前最佳方法有了进一步提升,并在双手数据集InterHand2.6M上验证了动态特征选择的有效性。最后,本文基于AR交互建立了手部遮挡数据集(Hand Occlusion Dataset based AR,HODA),以AR交互方式实现了无实物操作引导,这在机器人模仿学习中非常重要。此外,HODA采用文本引导的扩散模型生成自然且连贯的交互背景,并使用分类方法自动标注遮挡标签,利用泛化性、相似性实验以及不同数据量扩充数据集实验验证了HODA有效性。总之,本文主要贡献如下:

(1)面向复杂交互场景,提出了一种通用的遮挡关节与可见关节的分类方法,能准确识别单手、双手和手－物交互下的关节遮挡状态,并为三种场景下的7个公开数据集生成遮挡标签。

(2)设计了一个动态特征选择模块,提出了融合关节遮挡信息的手部姿态估计网络,高效利用可见关节信息,抑制遮挡关节影响,在公开数据集DexYCB、HO3D和HanCo上进行验证,估计精度比目前最佳方法有了进一步提升。

(3)基于AR交互提出了一个手部遮挡数据集HODA,利用扩散模型生成自然且连贯的交互背景,采用分类方法自动标注遮挡标签,利用泛化性、相似性以及扩充数据集实验验证了HODA的有效性。

2　相关工作

2.1　手部模型研究

早期的手部建模主要依赖几何基元,如高斯模型[27]、球体模型[28]和体素化模型[29]等,但由于几何基元缺乏细节描述,仅能用于手部粗糙表征[30]。具有线性混合蒙皮的三角网格模型采用主成分分析可以更好完成手部局部信息柔顺连接,但受到固定网格影响,其局部精细化能力不足[31]。故De等人[32]提出了一个三角网格可自适应的手部模型,通过引入缩放参数来适配多种手形,但由于其没有考虑物理运动合理性,手指接触时会出现穿模现象。MANO模型[33]同时考虑了手部局部信息和物理运动合理性,分别利用姿态参数和形状参数回归手部模型,受到研究者广泛关注,也被用于本文实现手部重建。

2.2　遮挡下的手部姿态估计研究

基于视觉的手部姿态自然、直观地揭示了人类交互意图,但手部交互时不可避免存在遮挡现象,造成手部关键信息丢失,严重影响手部全局估计精度,受遮挡问题的限制,专注于解决遮挡问题的研究成果很少。多视角和时序信息是有望增强遮挡域特征的有效手段。但前者冗余信息多、计算复杂[18],且受限于交互场景难以在日常交互中推广[5,34]。后者虽然不受交互场景限制,但是手部姿态估计误差会随着时间帧传递[22],造成估计误差累积[35]。Zhang等人[36]基于仿真合成图像解决真实图像由于遮挡引起的估计精度低问题,但合成图像与真实图像的差异依然没有得到解决。JoonKyu等人[24]将图像中可见区域作为主要特征,而遮挡区域作为次要特征,通过特征注入来提高手部姿态估计精度,但是随着次要特征占比增加,精度提升效果显著下降。Asuka等人[15]将相邻三根手指划分为一个整体,仅抑制遮挡关节对手部全局精度影响。Qi等人[37]采用多任务机制分别估计遮挡和可见关节,但难以保障遮挡分类精度。Xu等人[25]以手指为基本单元划分遮挡区域,指导帧间手指信息融合,但是难以真实反映局部遮挡信息,故提升效果有限。因此,提高关节遮挡分类精度,利用局部遮挡信息提高手部姿态估计精度有望解决遮挡难题。

2.3　手部数据集研究

手部数据集是影响手部姿态估计效果的重要因素之一,根据交互场景需求,可以将手部数据集划分为单手交互、双手交互和手－物交互。单手交互是最早开展研究的方向之一,围绕单手姿态估计的手部数据集也最丰富。NYU[35]、ICVL[38]、MSRA[39]和FPHA[40]等数据集使用了深度图并提供了3D关节标签。RHD[41]、STB[42]、FreiHAND[43]和Han-

Co[44]等数据集引入了像素级标签为网络提供了更直接的监督，缓解了从 RGB 图像回归 3D 手部姿态的病态问题[36]。随着单手姿态估计成果的涌现，双手和手－物交互也逐渐引起了学者的重视[45]，相关数据集也随之出现。InterHand2.6M[46]构建了双手交互数据集。DexYCB[47]，HO3D[48]是围绕手－物交互的数据集，包含了 3D 关节，MANO 标签。综上，虽然关于单手、双手和手－物数据集都已经取得重要进展，但是几乎没有数据集专注于遮挡问题，这也是制约遮挡下手部姿态精准估计的主要因素之一。最近，扩散模型在生成图像方面展现出巨大潜能[49]，也有学者通过扩散模型生成手部图像来丰富手部数据集，提高手部姿态估计效果[50-52]。因此，本文提出 AR 交互下含遮挡标签的手部数据集，并通过 ChatGPT 和文本引导的扩散模型丰富了数据集背景，有望促进遮挡下手部姿态估计研究。典型的手部数据集和 HODA 的详细对比信息如表 1 所示。

表 1 HODA 与手部公开数据集的比较

数据集	交互场景	注释方法	模态	分辨率	数量	网格	遮挡标签
NYU[35]	one-hand	computational	Depth	640×480	243K	×	×
ICVL[38]	one-hand	computational	Depth	320×240	18K	×	×
MSRA[39]	one-hand	computational	Depth	640×480	76K	×	×
FPHA[40]	one-hand	marker	RGB-D	1920×1080	105K	×	×
RHD[41]	one-hand	synthetic	RGB-D	320×320	44K	×	×
STB[42]	one-hand	manual	RGB-D	640×480	36K	×	×
FreiHAND[43]	one-hand	computational	RGB	224×224	134K	√	×
HanCo[44]	one-hand	computational	RGB	224×224	810K	√	×
InterHand2.6M[46]	two-hand	computational	RGB	512×334	2.6M	√	×
DexYCB[47]	hand-object	manual	RGB-D	640×480	582K	√	×
HO3D[48]	hand-object	computational	RGB-D	640×480	78K	√	×
OakInk[45]	hand-object	manual	RGB-D	848×480	230K	√	×
HODA(ours)	one-hand	computational	RGB-D	640×480	528K	√	√

3 预备知识

本节主要介绍分类方法的预备知识，包括 MANO 模型、手部结构先验知识和手部模型分割。MANO 模型是分类方法的基础，手部结构先验知识和手部模型分割是分类方法的前提。

3.1 MANO 模型

MANO 模型[33]同时考虑了手部局部信息和物理运动合理性，受到研究者广泛关注。MANO 模型通过分别回归姿态参数 $\theta \in \mathbb{R}^{48}$ 和形状参数 $\beta \in \mathbb{R}^{10}$ 来构建，由 778 个网格点 $\boldsymbol{V}_{3D} \in \mathbb{R}^{778\times 3}$ 和 21 个关节 $\boldsymbol{J} \in \mathbb{R}^{21\times 3}$ 构成。其中，姿态参数决定手部运动形态，形状参数决定手部几何轮廓。

3.2 手部结构先验知识

手部是人类最灵活的执行器官之一，由 1 个腕关节和 20 个手指关节构成，如图 2 所示。腕关节表示为 $\boldsymbol{J}_0$，手指关节表示为 $\boldsymbol{J}_a^b$，a 为手指编号，取值为 $\{1,2,3,4,5\}$，依次对应拇指、食指、中指、无名指和小指，b 为手指上的指节编号，取值为 $\{1,2,3,4\}$，依次对应掌指关节、近侧指间关节、远侧指间关节和指尖。为了便于开展分类方法研究，针对目标关节定义了相邻关节和非相邻关节，如图 2 蓝色和绿色关节所示。手指的每一个关节依次作为目标关节，相邻关节是与当前目标关节处于同一根手指，且距离最近的关节，而非相邻关节是手指关节中除相邻关节之外的其余关节。

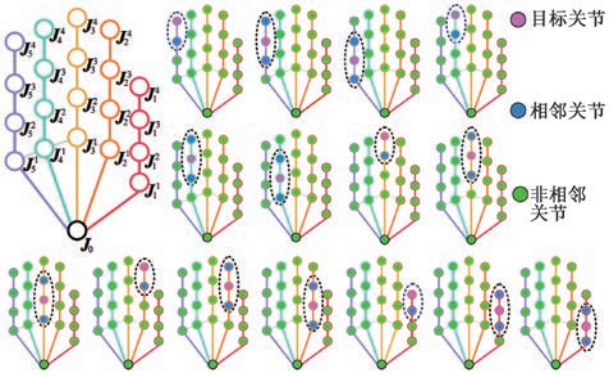


图 2 手部结构知识信息

3.3 手部模型分割

自遮挡是指手部关节被自身遮挡，由于手部是多自由度运动体，为了提高关节遮挡分类精度，将手部分割为 21 个独立运动单元（具有刚体运动属性[53]），运动单元由相邻手指关节之间的网格点集合构成。为了便于描述手部分割，基于 MANO 模

型将手部展成 T 型姿态,如图 3(a)所示。首先在 MANO 模型上建立手指垂直平面、手掌垂直平面和角平分面,然后基于网格点与三种平面的空间位置关系分割出 21 个独立运动单元。

手指垂直平面是指以相邻关节之间骨组织轴线为法向量,且过手指关节的平面。如图 3(b)所示,除指尖外,以 $\boldsymbol{J}_a^b$ 指向 $\boldsymbol{J}_a^{b+1}$ 作为法向量,并利用 $\boldsymbol{J}_a^b$ 建立手指垂直平面 Π_a^b ,表示如下:

$$\Pi_a^b=\{\boldsymbol{x}\in\mathbb{R}^3\mid(\boldsymbol{x}-\boldsymbol{J}_a^b)\cdot(\boldsymbol{J}_a^{b+1}-\boldsymbol{J}_a^b)=0\},\quad a\in\{1,2,3,4,5\},b\in\{1,2,3\}\tag{1}$$

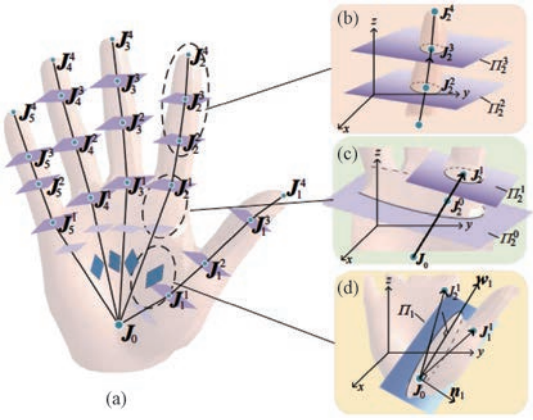


图 3　手部模型分割架构

利用公式 (1)可以建立 15 个手指垂直平面。手掌垂直平面是指以腕关节和掌指关节构成的矢量为法向量,且过两者连线三等分点的平面。如图 3(c)所示,由 $\boldsymbol{J}_0$ 指向 $\boldsymbol{J}_a^1$ 构建矢量作为平面法向量,并利用 $\boldsymbol{J}_0$ 和 $\boldsymbol{J}_a^1$ 连线靠近 $\boldsymbol{J}_a^1$ 的三等分点 $\boldsymbol{J}_a^0$ 建立手掌垂直平面 Π_a^0 ,表示如下:

$$\Pi_a^0=\{\boldsymbol{x}\in\mathbb{R}^3\mid(\boldsymbol{x}-\boldsymbol{J}_a^0)\cdot(\boldsymbol{J}_a^1-\boldsymbol{J}_0)\},\quad a\in\{1,2,3,4,5\}\tag{2}$$

利用公式 (2)可以建立 5 个手掌垂直平面。角平分面满足过相邻手指角平分线矢量,且垂直于相邻手指轴线所在平面。如图 3(d)所示,由 $\boldsymbol{J}_0$ 指向 $\boldsymbol{J}_a^1$ 和 $\boldsymbol{J}_0$ 指向 $\boldsymbol{J}_{a+1}^1$ 分别构建相邻手指轴线,进而得到角平分线矢量 $\boldsymbol{\omega}_j$,利用两相邻手指轴线与 $\boldsymbol{\omega}_j$ 进行叉乘得到矢量 $\boldsymbol{n}_j$ 作为平面法向量,并利用腕关节 $\boldsymbol{J}_0$ 建立角平分面 Π_j ,表示如下:

$$\boldsymbol{\omega}_j=\frac{\boldsymbol{J}_{a+1}^1-\boldsymbol{J}_0}{|\boldsymbol{J}_{a+1}^1-\boldsymbol{J}_0|}+\frac{\boldsymbol{J}_a^1-\boldsymbol{J}_0}{|\boldsymbol{J}_a^1-\boldsymbol{J}_0|}$$
$$\boldsymbol{n}_j=(\boldsymbol{J}_{a+1}^1-\boldsymbol{J}_0)\times(\boldsymbol{J}_a^1-\boldsymbol{J}_0)\times\boldsymbol{\omega}_j,a=j\tag{3}$$
$$\Pi_j=\{\boldsymbol{x}\in\mathbb{R}^3\mid(\boldsymbol{x}-\boldsymbol{J}_0)\cdot\boldsymbol{n}_j=0\},\quad j\in\{1,2,3,4\}$$

利用公式 (3)可以得到 4 个角平分面。最终得到手部分割后的 24 个空间平面。

最后通过判断网格点与 24 个空间平面的位置关系(如点在平面左或右、上或下),来聚类网格点,得到 21 个独立运动单元。位置关系计算为:

$$V_a^b(+)=\{\boldsymbol{v}\in\{\boldsymbol{V}_{3D}\}\mid(\boldsymbol{v}-\boldsymbol{J}_a^b)\cdot(\boldsymbol{J}_a^{b+1}-\boldsymbol{J}_a^b)>0\},\quad a\in\{1,2,3,4,5\},b\in\{0,1,2,3\}$$
$$V_a^b(-)=\{\boldsymbol{v}\in\{\boldsymbol{V}_{3D}\}\mid(\boldsymbol{v}-\boldsymbol{J}_a^b)\cdot(\boldsymbol{J}_a^{b+1}-\boldsymbol{J}_a^b)<0\},\quad a\in\{1,2,3,4,5\},b\in\{0,1,2,3\}$$
$$V_j(+)=\{\boldsymbol{v}\in\{\boldsymbol{V}_{3D}\}\mid(\boldsymbol{v}-\boldsymbol{J}_0)\cdot\boldsymbol{n}_j>0\},\quad j\in\{1,2,3,4\}$$
$$V_j(-)=\{\boldsymbol{v}\in\{\boldsymbol{V}_{3D}\}\mid(\boldsymbol{v}-\boldsymbol{J}_0)\cdot\boldsymbol{n}_j<0\},\quad j\in\{1,2,3,4\}\tag{4}$$

其中,$\boldsymbol{v}$ 为一个网格点,$V_a^b(+)$ 和 $V_a^b(-)$ 为位于垂直平面 Π_a^b 正方向或者负方向的网格点集合。$V_j(+)$ 和 $V_j(-)$ 为位于角平分面 Π_j 正方向或者负方向的网格点集合。基于网格点与所有平面的空间关系,对网格点进行聚类。计算方法为:

$$\begin{aligned}
&S_1^3=V_1^3(+)\cap V_1(+)\\
&S_a^3=V_a^3(+)\cap V_j(+)\cap V_{j-1}(-),\quad a=j\in\{2,3,4\}\\
&S_5^3=V_5^3(+)\cap V_4(-)\\
&S_1^b=V_1^b(+)\cap V_1^{b+1}(-)\cap V_1(+),b\in\{0,1,2\}\\
&S_a^b=V_a^b(+)\cap V_a^{b+1}(-)\cap V_j(+)\cap V_{j-1}(-),\quad a=j\in\{2,3,4\},b\in\{0,1,2\}\\
&S_5^b=V_5^b(+)\cap V_5^{b+1}(-)\cap V_4(-),b\in\{0,1,2\}\\
&S_0=\{\boldsymbol{V}_{3D}\}-\bigcup_{a=1}^{5}\bigcup_{b=0}^{3}S_a^b
\end{aligned}\tag{5}$$

其中,S 为 21 个独立的运动单元集合。需要强调,上述手部模型分割方法适用手部任何姿态,不局限于 T 型姿态。

4　遮挡分类方法

本节提出单手自遮挡、双手互遮挡和物体遮挡分类方法,如图 1 所示。分类方法由 MANO 重建、手部模型分割和遮挡判别器组成。MANO 重建可以由现有成果得到,手部模型分割在 3.3 节已经给出,故本节重点阐述遮挡判别器。

4.1　单手自遮挡分类

自遮挡是指手部关节被自身遮挡,进一步可以表述为手部关节被分割后的 21 个运动单元遮挡。

令目标关节所在运动单元为相邻运动单元，相邻运动单元遮挡发生在相邻运动单元轴线与相机光轴近似平行时，此时相邻运动单元可观测信息极少，投影策略难以准确辨别目标关节是否被相邻运动单元遮挡，而其余运动单元则无此现象。因此，如图 1(a)所示，相邻运动单元和非相邻运动单元分别采用骨眼角策略和投影策略建立遮挡判别器。为了提高分类效率，只要目标关节被相邻或非相邻运动单元遮挡，则无需再进行判别。

4.1.1 相邻运动单元遮挡判别器

在判别相邻运动单元遮挡时，骨眼角 φ 是指射线 L_1 和射线 L_2 的夹角。其中 L_1 是由相机原点指向目标关节，即相机视线，L_2 是由目标关节指向目标相邻关节，即骨组织轴线，如图 4 所示。当手部参照相机运动时，根据相机成像原理，φ 能够在相机空间内准确分类处于任何位置的相邻运动单元遮挡，不易受距离变化影响。在本文中，若 φ 小于 15°，则认为关节存在遮挡。

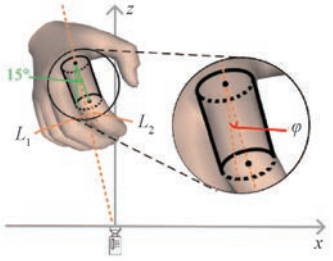


图 4 骨眼角策略

4.1.2 非相邻运动单元遮挡判别器

当目标关节被非相邻运动单元遮挡时需要同时满足深度条件和投影条件。深度条件是指目标关节深度值要大于非相邻运动单元深度值，前者深度值直接从 3D 坐标获取，后者深度值由非相邻运动单元上所有网格点深度均值确定。投影条件是指目标关节和非相邻运动单元基于相机参数投影到 2D 像素平面后，2D 目标关节 $\boldsymbol{J}_{2D}$ 应位于 2D 运动单元内，如图 5 所示。为了判别投影条件，利用 Melkman 算法[54]搜索 2D 运动单元得到双端队列 Q，对 Q 采用分段线性插值获得凸包闭合区域 O，O 是运动单元的最大包络区域，可以有效解决由 MANO 模型的表征能力有限导致的遮挡标签构建不准确问题，所以投影条件可以表示为：

$$Occ(\boldsymbol{J})=\begin{cases}\text{occlusion} & ;\boldsymbol{J}_{2D}\in O\\ \text{visible} & ;\boldsymbol{J}_{2D}\notin O\end{cases} \tag{6}$$

此外，分类方法还可以指明造成遮挡现象的具体运动单元，构建文本提示，潜在增强遮挡特征。

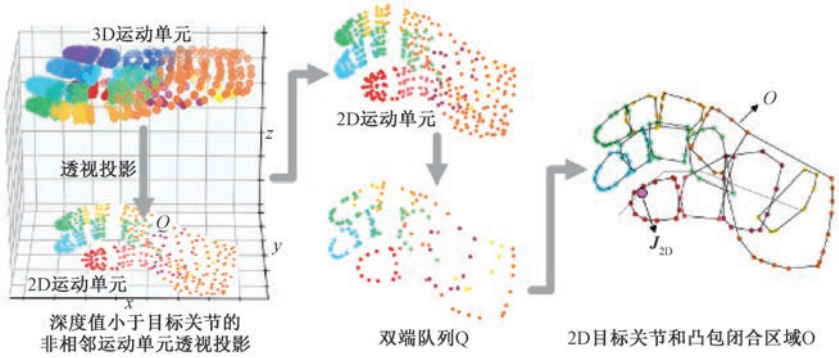


图 5 投影策略

4.2 双手互遮挡分类

双手互遮挡是指手部关节除了自遮挡，还可能会被另一只手遮挡。利用手部分割模块将双手 MANO 模型分割为 42 个运动单元。此时，双手可以视作拥有 42 个关节的“超级单手”。因此，自遮挡判别器可以直接迁移到双手互遮挡分类中，如图 1(b)所示。

4.3 物体遮挡分类

物体遮挡是指手部关节被所交互的物体遮挡。物体遮挡分类如图 1(c)所示，以公开数据集 DexYCB[47]和 HO3D[48]为例，首先获取手部 MANO 模型、物体点云和物体 6D 姿态，将物体点云视为一个非相邻运动单元，此时物体遮挡分类方法与自遮挡非相邻运动单元判别器思路相似。其中投影条件与自遮挡一致，但由于不同物体的形状和尺寸均各异，运动单元深度值不能简单由形心深度值代替，而是应选择物体上与目标关节最近的局部区域深度最小值来描述。本文以相机原点到目标关节连线为轴，以手指近似厚度(1 厘米)为直径构建圆柱体，取物体与圆柱体交集内点的深度最小值作为运动单元深度值。

5　融合关节遮挡信息的手部姿态估计网络

为了进一步提高遮挡下手部姿态估计精度,本文设计了动态特征选择模块,尝试在重要研究成果POEM[18]上融合遮挡与可见关节信息,如图 6 所示。网络由特征提取、动态特征选择和特征利用模块级联实现。

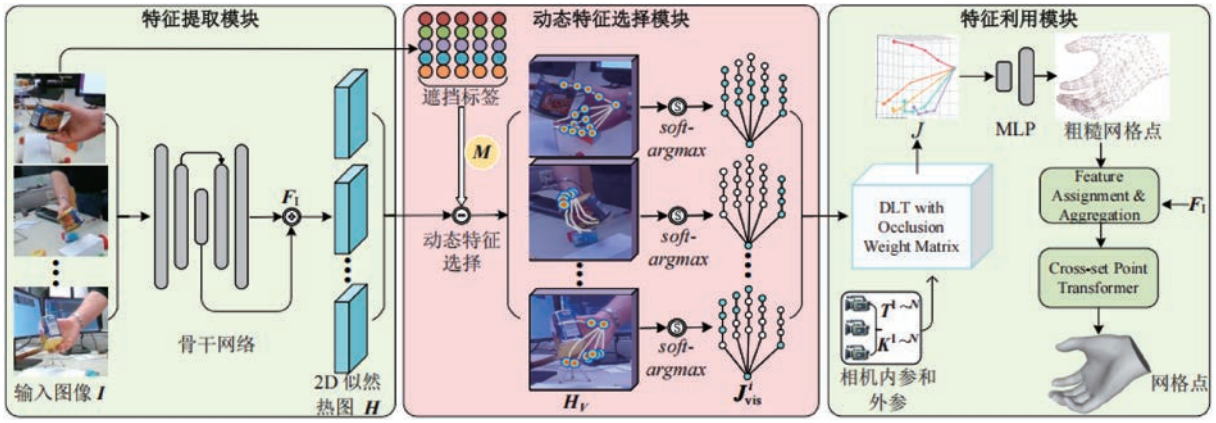


图 6　融合关节遮挡信息的手部姿态估计网络框架

5.1　特征提取模块

特征提取由 ResNet-50 和沙漏模块组成的主干网络完成,从 N 视角手部图像 $\boldsymbol{I}\in\mathbb{R}^{N\times256\times3}$ 中提取全局特征 $\boldsymbol{F}_{\mathrm{I}}\in\mathbb{R}^{N\times128\times32\times32}$,并得到 2D 似然热图 $\boldsymbol{H}\in\mathbb{R}^{N\times21\times32\times32}$。

5.2　动态特征选择模块

POEM 从多视角图像提取手部特征,认为每个视角信息对于获取 3D 关节位置同等重要,但使用全部信息进行估计不仅存在大量冗余,还未考虑遮挡造成精度下降的影响。本文认为可见关节对姿态估计有利,属于强相关特征,而遮挡关节对姿态估计不利,属于弱相关特征。为了充分利用强相关特征,本文利用关节遮挡状态构建遮挡权重矩阵进行动态特征选择,仅选择强相关特征进行利用,过滤掉弱相关特征。既避免了信息冗余,又避免了弱相关特征干扰。本文构建遮挡权重矩阵 $\boldsymbol{M}\in\mathbb{R}^{N\times21}$,其中 N 表示视角数量,21 为手部关节数量。$\boldsymbol{M}$ 中元素 (i,j) 取值为 1 或 0,其中 1 表示第 i 个视角中第 j 个关节可见,0 表示第 i 个视角中第 j 个关节被遮挡。如图 7 所示,利用权重矩阵 $\boldsymbol{M}$ 对每个视角的 21 个热图集合 $\boldsymbol{H}=\{h_1,h_2,\cdots,h_{21}\}$ 进行筛选,仅保留可见点的热图信息,并将可见点热图信息整合成一组新的热图集合,以此来抑制遮挡关节信息随着网络层数递增而向后传递,从而减少遮挡关节对手部姿态估计精度的影响,上述过程可表示为:

$$\boldsymbol{H}_{\mathrm{v}}=\boldsymbol{M}\circ\boldsymbol{H} \tag{7}$$

其中,$\boldsymbol{H}_{\mathrm{v}}$ 为强相关特征,$\circ$ 为哈达玛积。

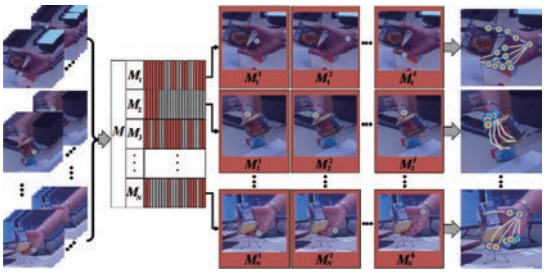


图 7　融合关节遮挡信息的手部姿态估计网络框架

5.3　特征利用模块

考虑到关节和网格点的稀疏差异性,本文采用串联多任务机制依次回归关节和网格点空间位置。

5.3.1　关节点回归

三角测量是计算机视觉中基于多视角的经典定位算法,故本文采用三角测量法中的直接线性变换(Direct Linear Transformation)回归关节空间位置,可以表示为:

$$\begin{aligned}\boldsymbol{J}_{\mathrm{vis}}^{1\sim N}&=soft-argmax(\boldsymbol{H}_{\mathrm{v}})\\ \boldsymbol{J}&=DLT(\boldsymbol{J}_{\mathrm{vis}}^{1\sim N},\boldsymbol{K}^{1\sim N},\boldsymbol{T}^{1\sim N})\end{aligned} \tag{8}$$

其中,$\boldsymbol{J}_{\mathrm{vis}}^{i}$ 是第 i 个视角的 2D 可见关节,$soft\text{-}argmax$ 通过指数归一化函数探索 $\boldsymbol{H}_{\mathrm{v}}$ 最大索引值的函数,$\boldsymbol{K}^{i}$ 和 $\boldsymbol{T}^{i}$ 是第 i 个视角下相机的内参和外参。

5.3.2　网格点回归

参考 POEM 特征利用,以回归的关节点 $\boldsymbol{J}$ 为输入,利用全连接回归粗糙网格点云。然后将多视角相机空间交集离散化为体素,在每个体素内放置一

个点,得到空间点云集合。最后使用 Cross-set Point Transformer 融合两组点云回归网格点精确位置。具体过程可参考文献[18]。

6 HODA 数据集

本文基于 AR 交互建立了手部遮挡数据集 HODA,以 AR 交互方式抓取虚拟物体不仅能实时反馈真实操作状态,还能通过无实物交互操作避免物体遮挡,这对于机器人模仿学习具有重要意义[55]。HODA 包含 528,000 张图像,采用分类方法自动标注自遮挡标签,并通过文本引导的扩散模型生成自然且连贯的交互背景。

6.1 环境准备

HODA 封闭数据采集平台由 Microsoft HoloLens、3 台 Intel RealSense D435 相机、绿幕和补光系统组成,如图 8 所示。实验选择 21 个 YCB 物体[56]作为操作对象,采集数据前,充分考虑人类操作习惯和物体形状,为每种物体设计 1—3 种抓取姿态。设计抓取姿态时,优先考虑上方、右侧和前方手势,若无法以舒适姿态完成抓取,则舍弃这种姿态,如图 9 所示。如布丁盒子满足上方、右侧和前方抓取手势,则有三种抓取姿态,而电钻仅能以右侧手势完成舒适抓取,所以只有 1 种抓取姿态。

(a) 数据采集平台的内部 (b) 数据采集平台的外部

图 8 数据采集平台

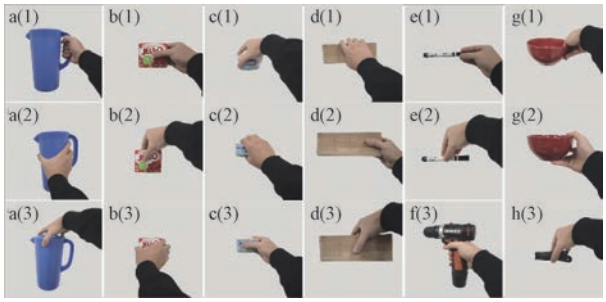


图 9 部分 YCB 物体的抓取手势

采集数据时依次将虚拟 YCB 物体模型放置在封闭平台中心区域,并可以实现手—物抓取交互。此外,3 台相机分别放置在左、右、前视角,同步采集数据。

6.2 数据收集

数据采集分为预抓取和抓取运动两个过程。预抓取是指手部从初始位置开始运动,直到成功抓取到物体。抓取运动是指手部在抓取物体情况下,随物体一起运动。预抓取过程受物体初始姿态影响,初始姿态越多,预抓取数据越丰富。本文根据人机工程学,同时兼顾姿态丰富性与工作量,为每个虚拟物体在平台中心选择了 7 种不同初始姿态,包括竖直姿态、后倾 45°姿态、后倾 90°姿态、左倾 45°姿态、左倾 90°姿态、右倾 45°姿态和右倾 90°姿态。在数据采集中,共 21 名实验者分别参与完成一个物体的预抓取和抓取运动,如图 10 所示。最后,HODA 共收集 1,056 个序列,每个序列 500 张图像,共计 528,000 张图像。

(a) 预抓取 (b) 抓取运动

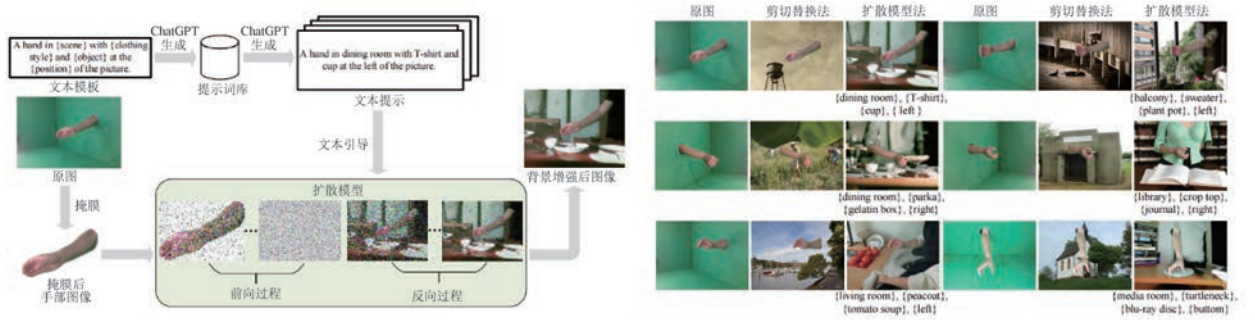
图 10 HODA 的数据收集过程

6.3 数据增强

自然、连贯的背景对于数据集至关重要,现有背景增强主要采用剪切替换,虽然能够增加背景的丰富性,但无法保障合理性。比如,手部出现在不合适位置,即位置不自然等[43];手部和背景无明显逻辑关系,即语义不连贯。扩散模型是一种典型的生成式模型,它能根据现有图像的内容、风格和语义,生成与图像相协调的新内容。本文采用了文本引导的扩散模型[57]为手部图像生成自然且连贯的背景。文本的质量直接影响背景的效果,场景、物体类型和物体位置约束可以促使文本满足连贯性要求,衣服的风格约束能够促使文本满足自然性要求。因此本文设计文本模版为“A hand in {scene} with {clothing style} and {object} at the {position} of the picture”。利用 ChatGPT 生成文本提示词库,如表 2 所示。然后,利用文本模板启发 ChatGPT 从中取材自动生成符合手部交互场景的文本提示。最后,将文本提示和掩膜后手部图像输入扩散模型[57],输出文本引导的背景增强图片,具体流程与结果如图 11 所示。

表 2　文本提示词库

场景		衣服样式		物体				物体在图像中位置
living room	kitchen	dress	shirt	remote control	cushion	knife	spoon	bottom
bedroom	bathroom	pants	skirt	pillow	alarm clock	toothbrush	soap	right
dining room	study	blouse	shorts	napkin	salt shaker	pen	notebook	left
guest room	office	jeans	T-shirt	lamp	hanger	stapler	paper clip	
library	sunroom	jacket	sweater	bookmark	journal	potted plant	magazine	
loft	nursery	coat	hoodie	box	quilt	pacifier	rattle cup	
game room	pantry	parka	scarf	dice	game controller	jar	tomato soup	
home theater	attic	suit	tie	popcorn	DVD	photo album	candle	
laundry room	hallway	shrug	peacoat	detergent	clothespin	umbrella	key rack	
lounge	workshop	turtleneck	cap	magazine	cup	screwdriver	nail	
basement	closet	hat	gloves	flashlight	toolbox	scarf	belt	
storage room	garage	bikini	pajamas	bin	tape	tire	wrench	
balcony	gym	robe	blazer	chair	plant pot	dumbbell	yoga mat	
wine cellar	playroom	kaftan	bolero	corkscrew	wine glass	toy car	puzzle	
media room	den	cardigan	jumpsuit	gelatin box	blu-ray disc	book	blanket	



(a) 文本引导的扩散模型背景增强流程　　(b) 使用剪贴替换法(左侧第2、5列)和扩散模型法(左侧第3、6列)增强背景示例

图 11　文本引导的扩散模型背景增强流程和结果

6.4　注　释

三角测量是计算目标物体空间位置的主要手段[58-59],也常被用于估计手部关节空间位置。本文先采用文献[44]的 2D 检测器获取每张图像 2D 关节位置,然后在每个视角中根据遮挡状态移除误差大的遮挡关节,再进行三角测量。移除原则为:关节在 3 个视角中均可见则都保留;在 1 个视角中遮挡则移除此视角;在 2 个视角中遮挡则移除与可见关节距离远的视角;在 3 个视角中遮挡则都保留。在得到 2D 关节后,使用三角测量[58]为数据集生成 3D 关节和网格点标签。部分数据集结果如图 12 所示。

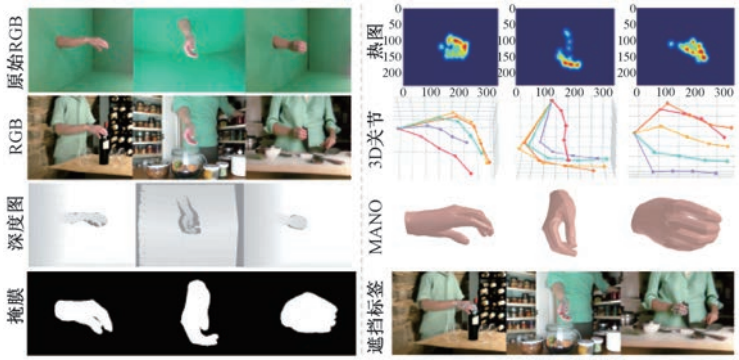


图 12　部分 HODA 结果

7　实　验

本节的 3 个实验分别回答以下问题:(1)遮挡分类方法是否有效?(2)遮挡信息对提高手部姿态估计精度是否有效?(3)HODA 数据集是否有效?

7.1　遮挡分类方法实验

据作者团队所知,目前还没有带遮挡标签的公开数据集,因此分类方法无法直接利用公开数据集

进行验证。对于少量数据，利用人工注释的准确性和一致性可以得到保障，故本文采用人工注释关节遮挡状态作为标签来验证分类方法有效性。

7.1.1 实验准备

本文面向单手交互、双手交互和手—物交互三种场景，选择了 7 个具有代表性的公开数据集和 HODA 进行分类实验。每个公开数据集通过等距抽样法得到 1 000 张图像，共 147 000 个手部关节。HODA 同样通过等距抽样法得到 2 000 张图像，共 42 000 个手部关节。实验招募 16 名志愿者，经过遮挡判断训练后，将 2 人划分为一组完成一个数据集关节遮挡注释。

7.1.2 实验结果

利用人工注释关节遮挡标签评价分类方法实验结果如表 3 所示。由于不同数据集差异性较大，所以实验结果也存在一定浮动，准确率最低为 95.07%，最高为 98.36%，精确率最低为 91.76%，最高为 99.21%，召回率最低为 87.65%，最高为 98.36%。HODA 是唯一专注于遮挡问题的数据集，其遮挡情况严重，但其各项指标依旧在所有单手数据集中表现优秀。其中，STB 数据集召回率偏低，这是因为其数据集中遮挡关节占比低。总体上，所有数据集实验结果均处于较高水平，证明了分类方法的有效性。分类方法还为 7 个公开数据集自动注释了遮挡标签，红色为遮挡关节，青色为可见关节，如图 13 所示。本文还分析了公开数据集中的遮挡情况，并给出了遮挡关节占全部关节的比例，如图 14 所示。

表 3 分类方法在不同数据集上的实验结果

数据集	交互场景	准确率(%)	精确率(%)	召回率(%)
HanCo[44]	one-hand	98.26	94.06	95.64
FreiHAND[43]	one-hand	98.19	97.42	94.76
RHD[41]	one-hand	97.85	94.03	96.63
STB[42]	one-hand	97.44	91.76	87.65
InterHand 2.6M[46]	two-hand	95.07	93.72	90.97
DexYCB[47]	hand-object	96.01	99.21	97.33
HO3D[48]	hand-object	98.26	97.07	98.36
HODA(ours)	one-hand	98.36	98.10	97.66

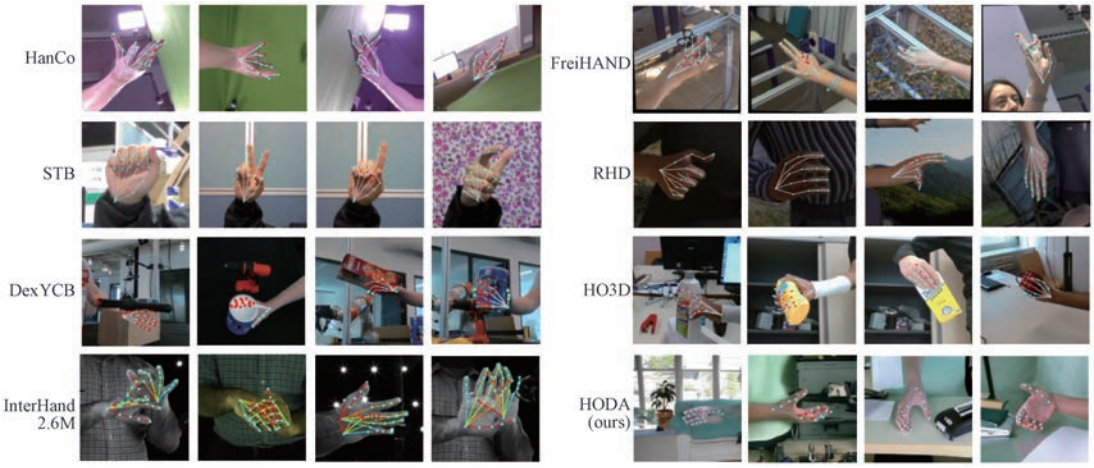


图 13 分类方法在不同数据集上的分类结果

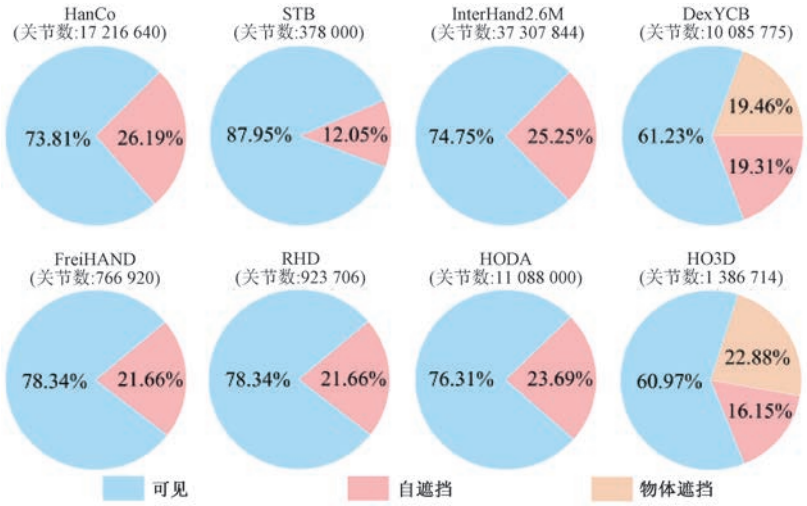


图 14 不同数据集中遮挡情况

7.2　遮挡关节数量与手部姿态估计精度实验

7.2.1　实验准备

为了验证遮挡关节数量对于手部姿态估计精度的影响,本文选择公开的单手数据集 FreiHAND、手物数据集 DexYCB 和本文构建的 HODA 数据集进行实验。依据遮挡分类方法对以上三个数据集测试集进行遮挡分类,对遮挡分类结果进行统计,如图 15 所示。根据统计结果,将测试集进一步分为无遮挡(0 个遮挡关节)、轻微遮挡(不大于 7 个遮挡关节)和重度遮挡(大于 7 个遮挡关节)。

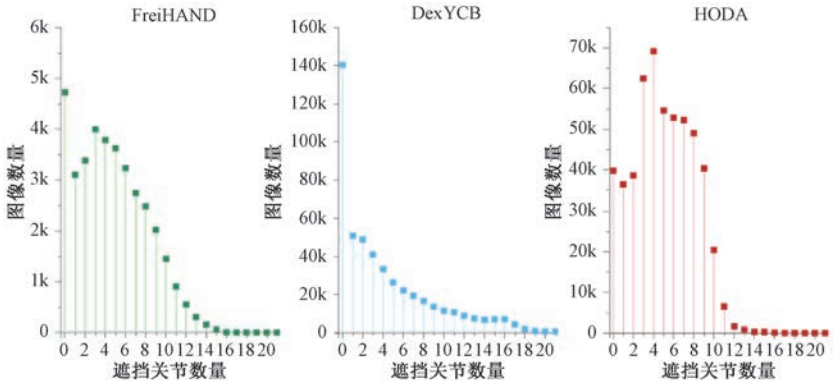


图 15　遮挡关节数量统计

7.2.2　实验结果

采用 AMVUR[72]、H2ONet[25] 和 Li 等人[74] 的方法依次对每个数据集中三个分类进行测试,用 PA-J-PE(普氏对齐后关节误差)作为评价指标,单位为 mm。实验结果如表 4 所示。每个数据集均是随着遮挡关节数量的增加,PA-J-PE 也明显增加。其中,各个数据集均是在重度遮挡情况下的关节预测精度最低,而在无遮挡情况下关节预测精度最高。实验结果表明,遮挡关节数量对关节精度会产生显著的不利影响。遮挡关节数量越多,对关节估计精度的影响越大。此外,本文还分析了不同遮挡类型对手部姿态估计的影响,如图 16 所示,各种遮挡类型都会导致精度的下降,因此,接下来我们对不同遮挡类型分别进行对比实验。

表 4　不同程度遮挡的 PA-J-PE 比较结果

遮挡程度	FreiHAND	DexYCB	HODA(ours)
无遮挡	6.77	5.22	4.85
轻微遮挡	9.06	6.29	5.11
重度遮挡	10.38	6.71	6.03

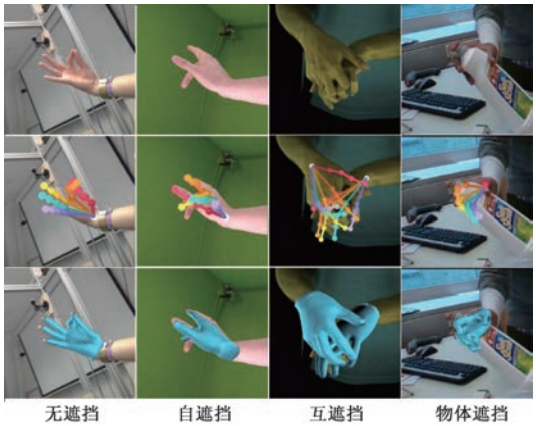


图 16　不同遮挡类型手部姿态估计效果图

7.3　融合关节遮挡信息的手部姿态估计实验

7.3.1　单手估计对比实验

(1)实验准备

本文选择单手数据集 DexYCB、HO3D 以及 HanCo 作为测试数据集。在 DexYCB 中,有 8 个视角图像,其中训练集包含 25 387 个多视角帧,测试集包含 4 951 个多视角帧。在 HO3D 中,有 5 个视角图像,考虑到多视角要求,本文选择“ABF1”、“BB1”、“GSF1”、“MDF1”和 “SiBF1”的序列作为训练集,“GPMF1”和“SB1”作为测试集。其中,训练集包含 9 087 个多视角帧,测试集包含 2 706 个多视角帧。在 HanCo 中,有 8 个视角图像,选择前 1 100 个序列作为训练集,后 418 个序列作为测试集。其中,训练集包含 79 851 个多视角帧,测试集包含 27 687 个多视角帧。

J-PE/V-PE、F-score 和 J-AUC/V-AUC 是手部姿态估计常用评态估计常用评价指标。J-PE/V-PE 表示关节或网格点空间位置误差,PA-J-PE/PA-V-PE 表示采用普氏对齐(Procrustes Alignment)后关节或网格点空间位置误差,单位为 mm。F-scores 表示关节预测值与真实值的召回率和精度的调和平均值,其中 F@5 和 F@15 表示阈值为 5mm 和 15mm。J-AUC/V-AUC 表示关节或网格点在误差范围内正确关键点曲线下的面积,在 DexYCB 上误差取值范围为 0～20mm,而 HO3D 上误差取值范围为 0～50mm。

实验在 Inter Xeon Gold 6248R CPU、NVIDIA

GeForce 3090 GPU 服务器上开展训练。Batch Size 设置为 32，每个实验都训练 100 轮，学习率初始值设置为 1×10^{-4}，70 轮后，学习率每 10 轮以 0.1 比例衰减，Cross-set Transformer 初始化均使用 xavier 均匀分布。此外，由于本文主要关注遮挡信息对手部姿态估计的积极作用，故在融合遮挡信息的手部姿态估计网络测试阶段，直接利用数据集中制作好的遮挡标签构建遮挡权重矩阵。

(2)实验结果

算法在手—物数据集 DexYCB、HO3D 和裸手数据集 HanCo 上实验结果如表 5 所示。本文提出的算法在 DexYCB、HanCo 数据集上的所有指标和阈值上达到了最佳表现，并且在 HO3D 数据集上的表现超越了大多数的算法。此外，本文提出的算法在 DexYCB 和 HO3D 数据集上相比于基准算法 POEM 也有明显提升，部分定性比较结果如图 17 所示。实验结果表明，通过动态特征选择模块融合关节遮挡信息可以有效提高手部姿态估计精度，不仅证明了抑制遮挡关节对可见关节精度提升的有效性，也证明了遮挡分类的必要性。另外，提出的算法在 DexYCB 数据集上速率为 17 帧/秒，在 HO3D 数据集为 21 帧/秒，在 HanCo 数据集为 18 帧/秒，其中每帧为一组多视角图像。

表 5 DexYCB、HO3D 和 HanCo 数据集上对比实验结果

	方法	J-PE ↓	V-PE ↓	PA-J-PE ↓	PA-V-PE ↓	F@5 ↑	F@15 ↑
DexYCB	METRO[60]	15.24	—	6.99	—	—	—
	Spurr[61]	17.34	—	6.99	—	—	—
	Liu[62]	15.27	—	6.58	—	—	—
	MobRecon[63]	14.20	13.1	6.40	5.6	0.508	0.921
	HandOccNet[24]	14.04	13.1	5.80	5.5	0.515	0.924
	H2ONet[25]	13.7	12.7	5.30	5.2	0.521	0.923
	Handbooster[64]	12.9	12.5	5.1	5.1	0.521	0.922
	HFL[65]	12.56	—	5.47	—	—	—
	IPNet[6]	8.03	—	—	—	—	—
	MVP[66]	6.23	9.77	4.26	8.14	—	—
	POEM[18]	6.06	6.13	3.93	4.00	—	—
	Ours	5.74	5.85	3.74	3.85	0.894	0.996
HO3D	Artib[67]	23.4	—	10.8	10.4	—	—
	IPNet[6]	19.3	—	8.8	8.4	—	—
	MVP[66]	18.72	20.95	10.44	10.04	—	—
	POEM[18]	17.27	17.2	9.6	9.97	—	—
	Ours	15.76	15.70	9.06	9.56	0.548	0.952
HanCo	Boukhayma et al.[68]	—	—	13.00	—	0.435	0.898
	Hasson et al.[69]	—	—	13.20	—	0.436	0.908
	Zimmermann et al.[43]	—	—	10.70	—	0.529	0.935
	Christian et al.[44]	—	—	10.20	—	0.548	0.943
	Kocabas et al.[70]	8.0	—	4.4	—	—	—
	Ours	6.73	6.78	3.44	3.57	0.901	0.996

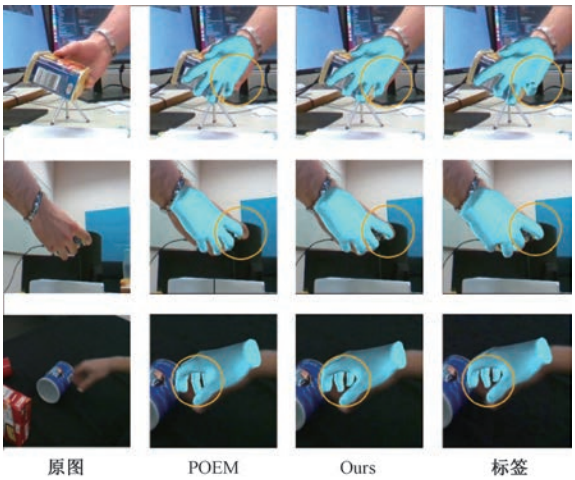


图 17 提出的算法与 POEM 的定性比较结果

7.3.2 双手估计对比实验

(1)实验准备

为了验证本文提出算法在双手数据集上效果，本文选择近三年、可训练双手估计算法 InterNet[46] 和 InterWild[70] 在最常用的双手公开数据集 InterHand2.6M 上开展对比实验研究。由于 POEM 是面向单手估计设计的网络框架，无法直接满足双手交互场景需求。因此，本文在不改变互遮挡状态下对 InterHand2.6M 数据集进行处理，具体过程为：在双手交互数据中，选择姿态差异大的 8 个相机视角作为输入，如“004”、“006”、“009”、“012”、“027”、“039”、“041”和“059”。在每个视角下，利用相机内、

外参数将左、右手 MANO 模型映射为 2D 信息,并划分得到 42 个独立运动单元。当左、右手独立运动单元出现重叠时,基于深度值来确定遮挡关系,并通过遮挡区域消除得到左、右手可见区域掩膜图片。如:当左手运动单元 i 和右手运动单元 j 出现重叠时,若 i 深度值大于 j 深度值,则左手运动单元 i 被右手运动单元 j 遮挡,在左手掩膜中消除与右手运动单元 j 重叠区域,得到左手可见域掩膜,反之亦然。最后,基于左手可见域掩膜得到右手图片数据,利用右手可见域掩膜得到左手图片数据,分别作为网络模型输入,对两组算法输出 J-PE 和 V-PE 取均值作为本文评价指标。

(2)实验结果

双手对比实验结果如表 6 所示,在 J-PE 指标上,InterNet 为 15.54 mm,InterWild 为 11.31 mm,精度远低于 POEM 的 9.02 mm 和本文提出算法的 7.87 mm。这是因为多视角可以提供更丰富的手部特征,输出更高的手部姿态估计精度。我们提出的算法估计精度高于 POEM,这是由于遮挡信息在我们提出算法中被最大程度抑制,使得手部特征更为准确。V-PE 指标有相似结果和相同结论,估计的可视化效果如图 18 所示。实验结果表明本文提出算法有望提高双手交互遮挡下手部姿态估计精度,同时进一步验证了遮挡分类标签的必要性。

表 6　处理后 InterHand2.6M 数据集上对比实验结果

方法	J-PE ↓	V-PE ↓
InterNet[46]	15.54 mm	15.88 mm
InterWild[71]	11.31 mm	11.66 mm
POEM	9.02 mm	9.26 mm
Ours	7.87 mm	8.13 mm

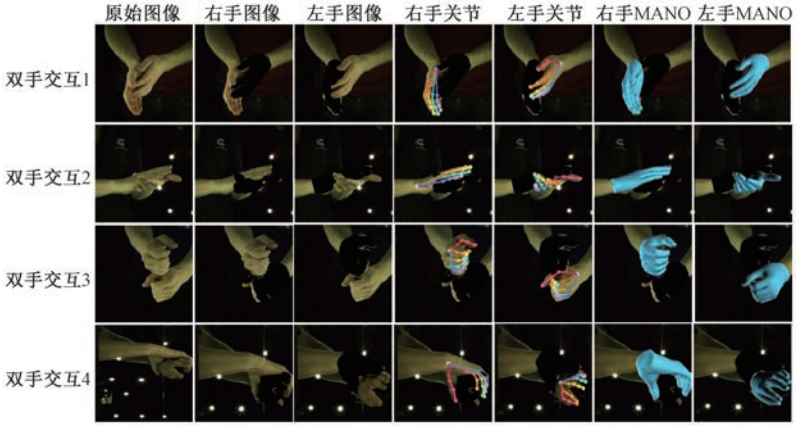


图 18　处理后双手数据集估计结果

7.3.3　权重矩阵消融实验

本节实验主要验证使用自遮挡和物体遮挡构建遮挡权重矩阵对精度的影响。因此,对仅使用自遮挡“Self”,仅使用物体遮挡“Obj”,和同时使用两种类型遮挡“All”开展实验研究,在 DexYCB 和 HO3D 数据集上的消融实验结果如表 7 所示,部分可视化结果如图 19 所示。在 DexYCB 数据集上,以对齐后 PA-J-PE 为例,利用遮挡信息构建遮挡权重矩阵的手部姿态估计精度均高于基线“Baseline”结果,Baseline 为 3.93mm,利用物体遮挡信息构建权重矩阵为 3.84mm,利用自遮挡信息构建权重矩阵为 3.79mm,利用物体遮挡和自遮挡信息构建权重矩阵为 3.74mm,这说明考虑遮挡信息对手部姿态估计精度提高有积极作用。但仅利用自遮挡或仅

表 7　权重矩阵消融实验结果　(单位:mm)

	方法	J-PE ↓	V-PE ↓	PA-J-PE ↓	PA-V-PE ↓	J-AUC ↑	V-AUC ↑
DexYCB	Baseline	6.06	6.13	3.93	4.00	0.68	0.70
	Obj	5.90	5.98	3.84	3.939	0.689	0.708
	Self	5.79	5.90	3.79	3.90	0.696	0.713
	All	5.74	5.85	3.74	3.85	0.698	0.716
HO3D	Baseline	17.27	17.2	9.6	9.97	0.63	0.66
	Obj	16.90	16.87	9.33	9.82	0.645	0.678
	Self	16.54	16.42	9.45	9.96	0.638	0.663
	All	15.76	15.72	9.06	9.56	0.660	0.688

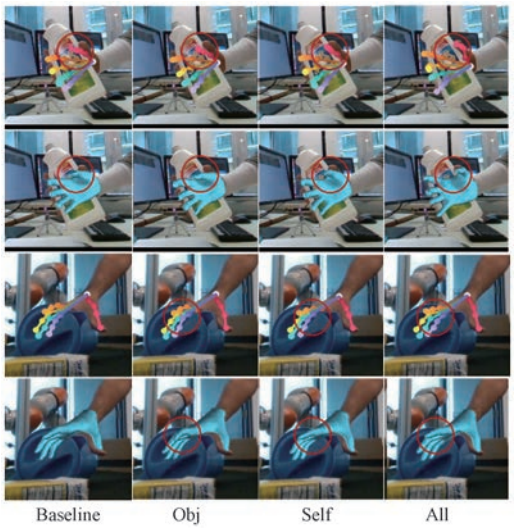


图 19 遮挡权重矩阵消融实验结果

利用物体遮挡对手部姿态估计精度提升并没有明显差异，分别为 3.79 mm 和 3.84 mm，这主要是由数据集中自遮挡和物体遮挡占比相似导致。此外，同时利用自遮挡和物体遮挡的手部姿态估计精度显著高于单独利用任何遮挡信息，为 3.74 mm，这说明在解决遮挡问题时，要综合考虑所有遮挡情况。HO3D 数据集具有相似结论。

为了进一步理解上述结果的深层次原因，本文尝试探讨了遮挡影响手部姿态估计精度的机理。在深度学习中，遮挡关节和可见关节的图像特征会从浅层传递到深层，网络层数越深，深层特征图上任一像素包含的信息越丰富，即可能同时包含多个遮挡和可见信息，最终影响手部可见关节估计精度。而引入遮挡权重矩阵是减少遮挡信息的深层次传递，抑制遮挡关节图像特征对最终可见关节估计精度的影响。无论是自遮挡还是物体遮挡，都只是图像遮挡的一部分，所以同时考虑两种遮挡信息时，可以最大程度提高手部姿态估计精度。

7.4 HODA 有效性分析

7.4.1 泛化性与相似性实验

(1)实验准备

实验在单手数据集 FreiHAND，双手数据集 InterHand2.6M，手—物数据集 DexYCB 和 HO3D 上进行测试。采用泛化性和相似性来验证 HODA 数据集的有效性。泛化性是指算法在原始数据集上完成训练，直接迁移到其他数据集上进行测试，其中“原始”是指提出算法的文章所采用的数据集。相似性是指多个算法在同一个数据集上进行训练和测试，比较估计精度。这里精度采用 PA-J-PE 表征。

(2)实验结果

在泛化性实验中，本文采用一些优秀的开源方法[72-74]作为手部姿态估计算法，在 FreiHAND、InterHand2.6M、DexYCB 和 HO3D 数据集上进行训练，然后迁移到 HODA 上进行测试，验证其泛化性，实验结果如表 8 所示。利用 FreiHAND 和 InterHand2.6M 数据集预训练模型在 HODA 上测试误差分别为 −1.4mm 和 1.3mm，没有明显差异。而采用 DexYCB 和 HO3D 数据集预训练模型在 HODA 上测试误差分别为 7mm 和 10.7mm，显著变大。事实上，虽然 HODA 是交互数据集，但随着 AR 的引入避免了物体遮挡问题，故 HODA 与 FreiHAND 和 InterHand2.6M 数据内容更接近，因此泛化性精度高。而 DexYCB 和 HO3D 含有大量物体遮挡情况，与 HODA 数据内容差异性大，故泛化性精度低[43]。

表 8 HODA 泛化性实验结果

训练 \ 测试	交互场景	原始	HODA (ours)
FreiHAND	one-hand	6.9	5.5
InterHand2.6M	two-hand	7.4	8.7
DexYCB	hand-object	5.6	12.6
HO3D	hand-object	8.7	19.4

在相似性实验中，本文同样采用 AMVUR[72]、H2ONet[25] 和 Li et al.[74] 算法依次在 DexYCB、HO3D 和 HODA 数据集上进行训练并测试，实验结果如图 20 所示。三种算法分别在 DexYCB 和 HO3D 上的测试精度差异较小。事实上，有效的数据集应该对多种算法具有相似精度。根据三种算法在 HODA 上的测试结果可知，精度相似性较好。综上，利用泛化性和相似性实验证明 HODA 是有效的。

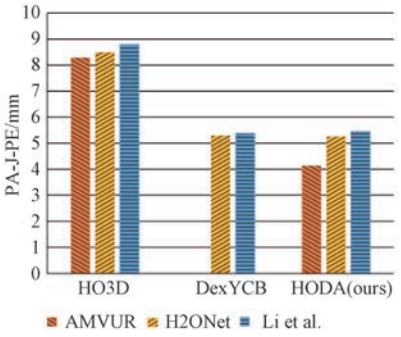


图 20 HODA 相似性实验结果

7.4.2 扩充数据集实验

(1)实验准备

对现有数据集的良好扩充是体现 HODA 有效

性的一种评价方式,本文选择相似性较高的单手FreiHAND数据集进行测试,从HODA中选择一定量数据扩充FreiHAND数据集,然后将CMR算法[75]在扩充后数据上进行训练,在FreiHAND原测试集上进行测试,选择PA-J-PE和PA-V-PE作为评价指标。由于FreiHAND训练集只有130,240张图像数据,而HODA训练集有408,000张图像数据,所以本文设计以下几个实验:(1)从HODA中随机选择两组各50,000张图像数据,分别扩充FreiHAND数据集;(2)从HODA中随机选择130,240张图像数据扩充FreiHAND数据集;(3)从HODA中选择全部408,000张图像数据扩充FreiHAND数据集。

(2)实验结果

实验结果如表9所示,随着扩充数据集中HODA数据量逐渐增加,手部姿态估计精度也随之提高,数据量从0张增加到408,000张,而PA-J-PE精度从7.50mm提升到7.29mm,PA-V-PE精度从7.60mm提升到7.40mm。这充分表明将HODA作为扩充数据可以显著提升在FreiHAND数据集上的估计效果。值得注意,HODA选择130,240张图像和408,000张图像精度提升幅度变小,这说明此时数据量已达饱和,完全可以满足算法训练需求,继续扩充数据,效果已经没有明显改善。

表9　扩充数据集实验结果

扩充数据量	J-PE ↓	PA-J-PE ↓
0	7.50	7.60
0,000	7.41	7.52
50,000	7.39	7.51
130,240	7.31	7.42
408,000	7.29	7.40

8　结　论

针对遮挡下手部姿态精准估计难题,本文提出面向复杂交互场景的单手、双手和物体遮挡分类方法,抑制遮挡域对可见域手部姿态估计精度的影响。设计了动态特征选择模块,在重要研究成果POEM上融合遮挡与可见标签,进而提出了串联特征提取、动态特征选择和特征利用的手部姿态估计网络,显著提高了遮挡下手部姿态估计精度。建立了基于AR交互的手部自遮挡数据集HODA,有利于开展机器人模仿学习研究。此外,为了丰富数据集,本文采用文本引导的扩散模型为手部图像生成连贯且自然的背景。在遮挡分类方法实验中,面向三种交互场景,选择了7个具有代表性的公开数据集和HODA进行分类实验,仅STB数据集的召回率略低于90%,值得注意专注于遮挡的HODA数据集在指标上均处于先进水平。在手部姿态估计实验中,根据遮挡关节数量将3个数据集划分成无遮挡,轻微遮挡以及重度遮挡,验证了随着遮挡程度的提升,估计精度逐渐下降;采用单手DexYCB、HO3D和HanCo作为测试数据集,与目前研究成果进行对比,本文提出算法融合自遮挡和物体遮挡信息,在所有指标和阈值上均显著高于基准算法POEM,在DexYCB、HO3D和HanCo数据集上指标达到了先进水平,同时通过在处理后双手数据集InterHand2.6M上进行测试,也证明了提出算法在解决双手互遮挡问题的有效性。通过权重矩阵消融实验明确得出,同时考虑自遮挡和物体遮挡估计精度优于单独利用任何一种遮挡。最后,在几个优秀的开源算法和4个公开数据集上验证了泛化性和相似性指标,并验证了HODA作为扩充数据能够提高其他数据集性能,共同证明了HODA数据集有效性。

本文提出遮挡分类方法,为7个公开数据集制作了遮挡标签,但是并没有充分利用遮挡标签来开展遮挡估计网络研究。后续可利用遮挡标签,针对遮挡估计网络展开研究。此外,本文融合遮挡信息提出动态特征选择模块,尝试在POEM网络上进行改进,但是动态特征选择模块受限于多视角并不具有通用性,难以对其他手部姿态估计网络给出融合遮挡信息的启发性思路。后续可探讨如何有效利用遮挡信息,提高手部姿态估计精度。

参　考　文　献

[1] Ge L, Ren Z, Yuan J. Point-to-point regression pointnet for 3d hand pose estimation//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 475-491

[2] Chen Z, Chen S, Schmid C, et al. gsdf: Geometry-driven signed distance functions for 3d hand-object reconstruction//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 12890-12900

[3] Ma Sheng-Lei, Li Jing-Hua, Kong De-Hui, et al. Hand 3D pose estimation based on two-branch multiscale attention. Chinese Journal of Computers, 2023, 46(7): 1383-1395. (in Chinese)
(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. 计算机学报, 2023, 46(7): 1383-1395.)

[4] Wang Y, Chen L L, Li J, et al. HandGCNFormer: A novel topology-aware ransformer network for 3d hand pose estima-

tion//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2023: 5675-5684

[5] Cheng J, Wan Y, Zuo D, et al. Efficient virtual view selection for 3d hand pose estimation//Proceedings of the AAAI Conference on Artificial Intelligence. Pittsburgh, USA, 2022, 36(1): 419-426

[6] Ren P, Chen Y, Hao J, et al. Two heads are better than one: Image-point cloud network for depth-based 3d hand pose estimation//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2023, 37(2): 2163-2171

[7] Myanganbayar B, Mata C, Dekel G, et al. Partially occluded hands: A challenging new dataset for single-image hand pose estimation//Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 2019: 85-98

[8] Meng H, Jin S, Liu W, et al. 3d interacting hand pose estimation by hand de-occlusion and removal//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 380-397

[9] Ren P, Wen C, Zheng X, et al. Decoupled iterative refinement framework for interacting hands reconstruction from a single rgb image//Proceedings of the IEEE/CVF International Conference on Computer Vision. Vancouver, Canada, 2023: 8014-8025

[10] Zhang S, Zheng T, Chen Z, et al. Ochid-fi: Occlusion-robust hand pose estimation in 3d via rf-vision//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 15112-15121

[11] Chen X, Wang G, Guo H, et al. Pose guided structured region ensemble network for cascaded hand pose estimation. Neurocomputing, 2020, 395: 138-149

[12] Miao Yong-Wei, Li Jia-Ying, Liu Jia-Zong, et al. Hand Gesture Recognition Based on Joint Rotation Feature and Fingert Distance Feature. Chinese Journal of Computers, 2020, 43(1): 78-92 (in Chinese)
(缪永伟, 李佳颖, 刘家宗等. 融合关节旋转特征和指尖距离特征的手势识别. 计算机学报, 2020, 43(1): 78-92)

[13] Patten T, Park K, Leitner M, et al. Object learning for 6D pose estimation and grasping from RGB-D videos of in-hand manipulation//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Prague, Czech Republic, 2021: 4831-4838

[14] Ghafoor M, Mahmood A. Quantification of occlusion handling capability of a 3D human pose estimation framework. IEEE Transactions on Multimedia, 2022, 25: 3311-3318

[15] Ishii A, Nakano G, Inoshita T. Occlusion-robust 3D hand pose estimation from a single RGB image//Proceedings of the International Conference on Machine Vision and Applications. Aichi, Japan, 2021: 1-5

[16] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA. 2016: 770-778

[17] Ren S, He K, Girshick R, et al. Object detection networks on convolutional feature maps. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(7): 1476-1481

[18] Yang L, Xu J, Zhong L, et al. Poem: Reconstructing hand in a point embedded multi-view stereo//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 21108-21117

[19] Zheng X, Wen C, Xue Z, et al. HaMuCo: Hand pose estimation via multiview collaborative self-supervised learning//Proceedings of the IEEE/CVF International Conference on Computer Vision. Vancouver, Canada, 2023: 20763-20773

[20] Tse T H E, Mueller F, Shen Z, et al. Spectral graphormer: Spectral graph-based transformer for egocentric two-hand reconstruction using multi-view color images//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 14666-14677

[21] Ren P, Sun H, Hao J, et al. Mining multi-view information: A strong self-supervised framework for depth-based 3D hand pose and mesh estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 20555-20565

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

[23] Chen J, Yan M, Zhang J, et al. Tracking and reconstructing hand object interactions from point cloud sequences in the wild//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2023, 37(1): 304-312

[24] Park J K, Oh Y, Moon G, et al. Handoccnet: Occlusion-robust 3d hand mesh estimation network//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 1496-1505

[25] Xu H, Wang T, Tang X, et al. H2onet: Hand-occlusion-and-orientation-aware network for real-time 3d hand mesh reconstruction//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 17048-17058

[26] Mahdikhanlou K, Ebrahimnezhad H. 3D hand pose estimation from a single RGB image by weighting the occlusion and classification. Pattern Recognition, 2023, 136: 109217

[27] Sridhar S, Mueller F, Oulasvirta A, et al. Fast and robust hand tracking using detection-Guided optimization//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3213-3221

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

[29] Malik J, Abdelaziz I, Elhayek A, et al. Handvoxnet: Deep voxel-based network for 3d hand shape and pose estimation from a single depth map//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

New Orleans, USA, 2020: 7113-7122

[30] De La Gorce M, Fleet D J, Paragios N. Model-based 3d hand pose estimation from monocular video. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(9): 1793-1805

[31] Schmidt T, Newcombe R A, Fox D. DART: Dense articulated real-time tracking//Proceedings of the Science and systems. Berkeley, USA, 2014, 2(1): 1-9

[32] Tzionas D, Ballan L, Srikantha A, et al. Capturing hands in action using discriminative salient points and physics simulation. International Journal of Computer Vision, 2016, 118: 172-193

[33] Romero J, Tzionas D, Black M J. Embodied hands: Modeling and capturing hands and bodies together. ACM Transaction on Graph,2017, 245:1-17

[34] Ge L, Liang H, Yuan J, et al. Robust 3d hand pose estimation in single depth images: From single-view cnn to multi-view cnns//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 3593-3601

[35] 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

[36] Zhang C, Jiao G, Di Y, et al. MOHO: Learning single-view hand-held object reconstruction with multi-view occlusion-aware supervision//Proceedings of the European Conference on Computer Vision. Seattle, USA, CVPR 2024: 9992-10002

[37] Ye Q, Kim T K. Occlusion-aware hand pose estimation using hierarchical mixture density network//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 801-817

[38] Tang D, Jin Chang H, Tejani A, et al. Latent regression forest: Structured estimation of 3d articulated hand posture//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 3786-3793

[39] Sun X, Wei Y, Liang S, et al. Cascaded hand pose regression//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 824-832

[40] Garcia-Hernando G, Yuan S, Baek S, et al. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 409-419

[41] Zimmermann C, Brox T. Learning to estimate 3d hand pose from single rgb images//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 4903-4911

[42] 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

[43] 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, Republic of Korea, 2019: 813-822

[44] Zimmermann C, Argus M, Brox T. Contrastive representation learning for hand shape estimation//Proceedings of the German Conference on Pattern Recognition. Bonn, Germany, 2021: 250-264

[45] Yang L, Li K, Zhan X, et al. Oakink: A large-scale knowledge repository for understanding hand-object interaction//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA 2022: 20953-20962

[46] Moon G, Yu S I, Wen H, et al. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 548-564

[47] Chao Y W, Yang W, Xiang Y, et al. DexYCB: A benchmark for capturing hand grasping of objects//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 9044-9053

[48] Hampali S, Rad M, Oberweger M, et al. Honnotate: A method for 3d annotation of hand and object poses//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2020: 3196-3206

[49] Bowen F, Gu W, Chenyangguang Z, et al. D-SCo: Dual-stream conditional diffusion for monocular hand-held object reconstruction//Proceedings of the European Conference on Computer Vision. Milan, Italy, 2024: 376-394

[50] Park J, Kong K, Suk-Ju K. AttentionHand: Text-driven controllable hand image generation for 3D hand reconstruction in the wild//Proceedings of the European Conference on Computer Vision. Milan, Italy, 2024: 329-345

[51] Zhang M, Fu Y, Ding Z, et al. Hoidiffusion: Generating realistic 3d hand-object interaction data//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA, 2024: 8521-8531

[52] Cha J, Kim J, Yoon J S, et al. Text2HOI: Text-guided 3D motion generation for hand-object interaction//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA, 2024: 1577-1585

[53] Gao D, Xiu Y, Li K, et al. DART: Articulated hand model with diverse accessories and rich textures. Advances in Neural Information Processing Systems, 2022, 35: 37055-37067

[54] Melkman A A, Shimony S E. Algorithms for parsimonious complete sets in directed graphs. Information Processing Letters, 1996, 59(6): 335-339

[55] Li Zhi, Pan Yue, Chen Dian-Sheng, et al. Fundus surgical behavior reproduction of robot based on imitation learning. Robot, 2024, 46(3): 361-369 (in Chinese)
(李至, 潘越, 陈殿生等. 基于模仿学习的眼底手术行为机器人复现. 机器人, 2024, 46(3): 361-369)

[56] Xiang Y, Schmidt T, Narayanan V, et al. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes//Proceedings of the Robotics: Science and Systems. Pittsburgh, USA, 2018:1-10

[57] Eshratifar A E, Soares J V B, Thadani K, et al. Salient object-aware background generation using text-guided diffusion models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA, 2024: 7489-7499

[58] Remelli E, Han S, Honari S, et al. Lightweight multi-view 3D pose estimation through camera-disentangled representation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2020: 6040-6049

[59] Iskakov K, Burkov E, Lempitsky V, et al. Learnable triangulation of human pose//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 7718-7727

[60] Lin K, Wang L, Liu Z. End-to-end human pose and mesh reconstruction with transformers//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 1954-1963

[61] Spurr A, Iqbal U, Molchanov P, et al. Weakly supervised 3d hand pose estimation via biomechanical constraints//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 211-228

[62] Liu S, Jiang H, Xu J, et al. Semi-supervised 3d hand-object poses estimation with interactions in time//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 14687-14697

[63] Chen X, Liu Y, Dong Y, et al. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 20544-20554

[64] Xu H, Li H, Wang Y, et al. HandBooster: Boosting 3D hand-mesh reconstruction by conditional synthesis and sampling of hand-object interactions//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA, 2024: 10159-10169

[65] Lin Z, Ding C, Yao H, et al. Harmonious feature learning for interactive hand-object pose estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 12989-12998

[66] Zhang J, Cai Y, Yan S, et al. Direct multi-view multi-person 3d pose estimation. Advances in Neural Information Processing Systems, 2021, 34: 13153-13164

[67] Yang L, Li K, Zhan X, et al. Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 2750-2760

[68] Boukhayma A, Bem R, Torr P H S. 3d hand shape and pose from images in the wild//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 10843-10852

[69] 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

[70] Kocabas M, Karagoz S, Akbas E. Self-supervised learning of 3d human pose using multi-view geometry//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 1077-1086

[71] Moon G. Bringing inputs to shared domains for 3D interacting hands recovery in the wild//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 17028-17037

[72] Jiang Z, Rahmani H, Black S, et al. A probabilistic attention model with occlusion-aware texture regression for 3D hand reconstruction from a single RGB image//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 758-767

[73] Yu Z, Huang S, Fang C, et al. Acr: Attention collaboration-based regressor for arbitrary two-hand reconstruction//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 12955-12964

[74] Shuang F, He W, Li S. 3D hand reconstruction via aggregating intra and inter graphs guided by prior knowledge for hand-object interaction scenario. Journal of Visual Communication and Image Representation, 2024, 100: 104129

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

LI Shao-Dong, Ph. D., assistant professor. His research interests include hand pose estimation and long-horizon robotics operations.

LUO Kai, M. S. candidate. His research interests include hand pose estimation.

HUANG Yuan-Zhi, undergraduate. His research interest is robotics.

LIU Xi, Ph. D. candidate. His research interest is hand pose estimation.

SHUANG Feng, Ph. D., professor. His research interests include computer vision, intelligent robot.

GAO Fang, Ph. D., professor. His research interests include computer vision, quantum control.

Background

3D hand pose estimation is a critical component in computer vision. Existing hand pose estimation methods are limited by the problem of missing features due to occlusion. However, most research doesn't adequately address the occlusion problem, while some methods have been proposed, their effect is still limited.

In this paper, we propose a joint occlusion classification method that can accurately classify self-occlusion, mutual occlusion, and object occlusion in complex interaction scenarios. We generate occlusion labels for public datasets. To verify the significance of the method, we propose a hand pose estimation network fusing joint occlusion information, which utilizes the joint occlusion status for dynamic feature selection, suppressing the accuracy degradation caused by occluded joint features. In addition, we present a hand occlusion dataset, created using augmented reality to simulate grasp interactions, with natural backgrounds enhanced via a text-guide diffusion model. Experimental results on eight datasets validate that the classification accuracy of the method exceeds 95.07%. And the hand pose estimation network fusing joint occlusion information achieves excellent performance on the DexYCB and HanCo datasets, it also outperforms most methods on the HO3D dataset. The proposed dataset is validated by generalizability, similarity, and complementary dataset experiments. Our code, dataset, and videos will be available on https://github.com/lishaodong-lab/HODA.

This work was supported in part by Natural Science Foundation of Guangxi under Grant 2023GXNSFBA026069, General Program of Guangxi Natural Science Foundation under Grant 2025GXNSFAA069931, and Innovation Project of Guangxi Graduate Education (No. YCSW2023056).