E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
©中国图象图形学报版权所有

中图法分类号:TP18; TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2026)02-0628-14

论文引用格式:Jia D, Wang J C, Han X F, Zhang F and Wang X. 2026. 3D hand pose estimation network integrating hand and object features. Journal of Image and Graphics, 31(2):0628-0641(贾迪, 王建淳, 韩雪峰, 张藩, 王骁. 2026. 融合手物特征的三维手部姿态估计网络. 中国图象图形学报, 31(2):0628-0641)[DOI:10.11834/jig.250270]

融合手物特征的三维手部姿态估计网络

贾迪[1,2],王建淳[2*],韩雪峰[1],张藩[2],王骁[2]

1. 辽宁工程技术大学鄂尔多斯研究院,鄂尔多斯　017000; 2. 辽宁工程技术大学电子与信息工程学院,葫芦岛　125105

摘　要:**目的**　手部姿态估计作为人机交互的核心感知技术,在复杂交互场景下,面临多尺度特征融合过程中信道丢失以及手物交互过程中遮挡干扰等挑战,现有方法多依赖单一特征提取策略或静态注意力机制,无法同时兼顾精度与鲁棒性。为此,提出一种融合手物特征的三维手部姿态估计网络(hand object collaborative enhancement network,HOCEN),旨在通过多层次跨模态交互优化,提升遮挡场景下的手部姿态估计性能。**方法**　设计双流特征金字塔,通过双向跨尺度信息聚合捕获局部细节与全局语义依赖,缓解传统特征金字塔的通道信息丢失问题;给出一种基于外部注意力的动态调整模块,对提取后的手部特征进行动态注意力权重分配,抑制噪声干扰;构建双流协同注意力机制,结合手—物几何约束与语义互补特性,增强跨模态特征对齐能力;通过层级特征解码器重构精准的手部姿态参数。**结果**　在Dex-YCB(dexterous-YCB)与HO3D(hand-object 3D)数据集上的实验结果表明,本文方法在遮挡场景下的手部关节定位精度高于当前的主流模型,在Dex-YCB数据集上,手部姿态估计指标MPJPE(mean per joint position error)和PA-MPJPE(procrustes-aligned mean per joint position error)分别达到12.4 mm和5.4 mm,均优于SemGCN(semantic graph convolutional network)、HFL(harmonious feature learning)等先进模型,在HO3D数据集上,手部姿态估计指标Join和Mesh上分别达到9.2 mm和9.1 mm,实现了极低的误差。此外,在Dex-YCB与HO3D数据集上分别进行消融实验,进一步证明各模块在手—物协同估计指标上的独立贡献与协同增益。**结论**　本文提出一种基于动态手物交互特征融合的三维手部姿态估计网络架构,通过跨模态特征协同建模机制有效提升姿态估计精度。实验结果表明,本文方法在复杂交互场景下具有较高的鲁棒性与泛化能力,提出的动态特征校准与手物协同策略为提升遮挡场景下的手部姿态估计提供了全新的解决方案。

关键词:手势姿态估计;双流金字塔特征融合;动态注意力机制;动态特征调整;手物协同增强

3D hand pose estimation network integrating hand and object features

Jia Di[1,2], Wang Jianchun[2*], Han Xuefeng[1], Zhang Fan[2], Wang Xiao[2]

1. Ordos Institute of Liaoning Technical University, Ordos　017000, China;

2. School of Electronic and Information Engineering, Liaoning Technical University, Huludao　125105, China

Abstract: **Objective** The estimation of hand poses is a crucial technology in the field of human-computer interaction, playing an important role in many complex interactive scenarios such as virtual reality, augmented reality, and robotics. How-

收稿日期:2025-07-01;**修回日期**:2025-09-03;**预印本日期**:2025-09-10

*通信作者:王建淳　lntu_wangjianchun@163.com

基金项目:国家自然科学基金项目(61601213);内蒙古自治区重点研发和成果转化计划(2025YFHH0047);辽宁省教育厅重点资助项目(LJ212410147003);辽宁工程技术大学鄂尔多斯研究院校地科技合作培育项目(YJY-XD-2023-003)

Supported by: National Natural Science Foundation of China(61601213); Science and Technology Plan Project of Inner Mongolia Autonomous Region, China(2025YFHH0047); Key Project of Liaoning Provincial Department of Education(LJ212410147003); University-Local Government Scientific and Technical Cooperation Cultivation Project of Ordos Institute-LNTU (YJY-XD-2023-003)

ever, owing to the high flexibility of hand movements, the complexity of self-occlusion, and the interaction between hands and objects, hand pose estimation remains a challenging task, especially in real-world scenarios. Traditional methods often struggle with issues such as insufficient multiscale feature fusion, channel loss, and occlusion interference. Most of the existing approaches tend to rely on either single-feature extraction strategies or static attention mechanisms, which frequently fail to achieve a balance between accuracy and robustness, particularly in occluded environments. As a result, hand pose estimation in these challenging settings often suffers from low accuracy. To address these challenges, this study proposes a novel 3D hand pose estimation network that integrates hand-object features, named hand-object collaborative enhancement network (HOCEN). The proposed method aims to enhance the robustness and accuracy of hand pose estimation in occlusion-prone environments by optimizing multilevel cross-modal interactions between hands and objects. **Method** First, the network introduces a dual-stream feature pyramid network (DS-FPN), aiming to capture local details and global semantic dependencies through bidirectional cross-scale information aggregation. This approach alleviates the common channel loss problem in traditional feature pyramid network (FPN), especially when dealing with multiscale features. Traditional FPNs usually fuse features through an upward path, which typically leads to the loss of fine details such as fingertip and joint textures, specifically in complex occlusion scenarios. By contrast, DS-FPN utilizes upward and downward information flows, enabling more excellent feature refinement. The downward flow enhances local details, while the upward flow injects global posture and semantic information. After these steps are completed, a second downward flow process is adopted to integrate the fine local details with global information, thereby forming an optimized feature fusion path. This path can enhance the low-level details and high-level global context of hand features, and this bidirectional interaction mechanism exists in the DS-FPN, significantly improving the accuracy and efficiency of hand feature extraction, especially in complex occlusion scenarios. Second, this study introduces a dynamic adjustment module (DAM) based on an external attention mechanism. This module dynamically adjusts the attention weights of the extracted hand features, ensuring that important features are emphasized while suppressing noise interference. The dynamic feature adjustment ensures that the hand features can effectively adapt to complex environments, thus achieving accurate feature extraction. The attention mechanism guided by external memory adapts to different conditions, optimizes feature calibration, and improves the overall robustness. Third, a hand feature enhancement module is constructed, aiming to enhance the feature alignment between hands and objects. This dual-stream collaborative attention mechanism combines the geometric constraints and semantic complementarity of hands and objects, thereby improving the cross-modal feature alignment between them. The HFE module fuses hand and object pose features through a multihead attention mechanism. This fusion enables the spatial information between the hands and objects to achieve precise alignment, which is crucial for tasks involving interactions between hands and objects. Finally, a hierarchical feature decoder is adopted, which is used to reconstruct 3D hand pose parameters. This decoder extracts the optimized features from the previous layers and accurately predicts a 3D hand pose on the basis of the estimated joint positions and pose parameters, thus reconstructing the 3D model of the hand. **Result** Extensive experimental validation is performed on the publicly available dexterous-YCB(Dex-YCB) and hand-object 3D (HO3D) datasets to evaluate the proposed method's effectiveness. Experimental results demonstrate that HOCEN outperforms existing state-of-the-art models in hand pose estimation accuracy, particularly in complex interaction and occlusion scenarios. On the Dex-YCB dataset, the proposed model achieves a mean per-joint position error (MPJPE) of 12.4 mm and a Procrustes-aligned MPJPE of 5.4 mm, which is superior to advanced models such as semantic graph convolutional network(SemGCN) and harmonious feature learning(HFL). Similarly, on the HO3D dataset, the model achieves very low joint and mesh errors of 9.2 mm and 9.1 mm, respectively, proving the model's accuracy in challenging conditions. Ablation studies further validate the individual contributions and synergistic effects of each module, with the DS-FPN module improving multiscale feature fusion, DAM enhancing feature robustness, and the HFE module strengthening hand-object interaction alignment. **Conclusion** This study proposes a 3D hand pose estimation deep network based on the fusion of dynamic hand-object interaction features. It effectively improves the hand pose estimation in occluded environments and solves key problems such as multiscale feature fusion, noise interference, and interference between hand and object features. Experimental results show that this method has strong robustness and generalization ability in complex interaction scenarios. Dynamic feature calibration and hand-object collaboration strategies provide a novel solution for solving the prob-

lem of hand pose estimation in occluded situations.

Key words: gesture estimation; double-flow pyramid feature fusion; dynamic attention mechanism; dynamic feature adjustment; hand and object synergistic enhancement

0 引 言

手势姿态估计作为计算机视觉领域重要研究方向，在虚拟现实、增强现实与机器人学习等应用场景中具有广泛应用价值。然而，由于手部的高度灵活性、复杂的自遮挡以及手部与物体交互时的遮挡问题，使得该任务仍然面临诸多挑战。近年来，基于单目RGB图像的手势估计方法取得了显著进展，尤其是手部—物体交互场景下的手势姿态估计已然成为当下的研究热点。

在过去的几年中，特征金字塔网络（feature pyramid network，FPN）在计算机视觉领域取得了显著进展，特别是在手势姿态估计方面。最初，传统方法主要依赖单尺度特征进行手势姿态估计，但这些方法在处理不同尺度的手势和复杂背景时存在局限性。为解决这些问题，Lin等人（2017）提出特征金字塔网络（FPN），通过构建自顶向下的路径和横向连接，将高层语义特征与低层细节特征相结合，从而在目标检测任务中取得了优异的性能。然而，FPN最初并未针对手势姿态估计任务进行设计。此外，Chen等人（2018）提出级联金字塔网络（cascaded pyramid network，CPN），利用不同层次的多尺度特征映射，从局部和全局特征中获得更多推断，并提出关键点挖掘损失来针对较难预测的关键点。该方法在多人姿态估计中表现出色，但在处理手部细节时依然存在问题。随后，针对场景复杂手部特征提取困难的问题，Lee等人（2019）使用FPN作为手部特征提取网络以获得多尺度特征图，同时使用新型残差网络作为模型特征提取部分的主干网络，以尽可能地提取手部特征。然而，在处理高自由度的手部动作时，模型的精度仍需提升。针对这一问题，He等人（2021）设计了一种全流双向融合网络，通过整合局部和全局互补信息进行优化表示，从而显著提高物体6D姿态估计的精度，但该方法主要针对物体姿态估计，未专门考虑手势姿态估计的特点。为了解决这一问题，贾迪等人（2023）提出一种针对单目视觉的手势姿态估计的多尺度特征融合网络，利用通道变换模块和全局回归模块的结合，提高对手部关节点位置的预测精度。然而，该方法在处理复杂的手部姿态和自我遮挡问题时仍存在挑战。

近年来，基于注意力机制的方法在手势姿态估计任务中取得了较大进展。最初，传统方法主要依赖卷积神经网络（convolutional neural network，CNN）处理手势姿态估计任务，但这些方法在捕捉长距离依赖和全局信息方面存在局限性。为了解决这一问题，Zheng等人（2021）提出PoseFormer，这是一种纯粹基于Transformer的方法，用于视频中的3D人体姿态估计，不涉及卷积架构。该方法设计了一个时空Transformer结构，以全面建模每帧内的人体关节关系以及帧间的时间相关性，从而能够精准预测中心帧的三维人体姿态。这项研究表明，Transformer可以有效替代CNN在姿态估计中的作用，尤其在时间建模方面具有显著优势。然而，PoseFormer在处理手部细节和复杂手势姿态时仍然存在一定挑战。在此基础上，Chen等人（2021）提出A2J-Transformer，以解决从单个RGB图像估计3D交互手部姿态的问题。他们将现有的A2J框架扩展到RGB输入，并结合局部—全局Transformer，使模型能够同时捕获手部的局部细节和全局关节信息。实验表明，该方法在多种手势数据集上取得了优异表现，尤其在复杂的手势交互场景下效果更加突出。然而，A2J-Transformer主要关注手部关节的局部信息，在整体估计手臂和手部动态关系时仍有一定局限性。为进一步提升手臂和手部的动态建模能力，Liu等人（2022）提出空间—时间并行手臂和手部运动Transformer（spatial-temporal parallel arm-hand motion transformer，PAHMT），用于从单目视频中估计手臂和手部运动状态。该方法采用并行的空间—时间Transformer模块，以联合建模手臂和手部之间的动态关系，并引入新的损失函数来增强预测结果的平滑性和准确性。通过这种方式，PAHMT在视频序列建模方面取得了更稳定的性能，使其能够更精准地预测手势的动态变化。然而，当面对复杂的自我遮挡和动作模糊问题时，该方法的表现仍然受到一定限制。针对自我遮挡和模糊动作识别的挑战，Wen等人

(2023)提出一种层次化时间Transformer,以改进第一人称视角RGB视频中的3D手部姿态估计和动作识别。他们观察到手部姿态估计和动作识别在时间粒度上的差异,同时又在语义上存在密切关联,因此设计了由两个级联Transformer编码器组成的层次化框架。第1个Transformer侧重于短期时间线索,以增强手部姿态估计的精度,而第2个Transformer则在更长时间跨度上聚合帧级别的姿态和物体信息,以提升动作识别的能力。实验结果表明,该方法在大规模第一人称数据集上达到了最优性能,并显著提高手势姿态估计的稳定性和鲁棒性。

无论是单独依赖FPN进行特征提取还是单独依赖Transformer进行特征增强,都不足以有效应对手部预测中的多尺度融合遮挡鲁棒性和时序建模等挑战。为了解决这些问题,研究者们提出结合FPN和Transformer的方法。Park等人(2022)提出一种HandOcc-Net网络,采用了特征金字塔和空间注意力机制,结合特征注入模块来处理遮挡问题,HandOccNet通过特征注入机制(feature injection mechanism)增强遮挡区域的特征表达能力。该方法设计了FIT(feature injecting Transformer)和SET(self-enhancing Transformer)两个模块,运用注意力机制学习遮挡区域与未遮挡区域的映射关系,并优化遮挡区域的特征,显著提高3D手部网格估计的精度,尤其是在复杂交互数据集上的表现。Lin等人(2023)提出HFL-Net(harmonious feature learning network),采用单流与双流骨干网络结合的方式,实现和谐的手—物体特征学习,并通过自注意力机制提升手部和物体特征之间的交互关系,从而提高手部姿态识别的准确度。尽管HandOccNet和HFL-Net均结合了FPN和Transformer的方法,但两种方法对严重遮挡情况无法实现手部细节重现。HandOccNet使用单流的主干网络同时提取手部和物体的特征时,会出现手部和物体特征之间的“竞争”,而HFL-Net虽然使用双流主干网络,将特征提取的1、2、5层对手物特征进行共同学习,3、4层分别对手、物特征进行独立学习,但其使用的手物特征提取网络仍然是传统的FPN,其采用简单卷积进行通道降维,会导致通道信息的丢失。

针对上述问题,本文方法结合FPN和Transformer,采取手物特征提取双流网络结构,提出一种融合手物特征的三维手部姿态估计网络(hand object collaborative enhancement network,HOCEN),该网络不仅可提高手部提取信息的有效性,而且还能极大程度地完成手部姿态估计的精度。

本文主要贡献如下:1)提出双流手部特征金字塔网络(dual-stream hand feature pyramid network,DS-FPN)。该模块采用双向多尺度特征融合机制,通过跨层级特征交互有效缓解局部信息缺失问题,对手部特征信息的提取具有更好的准确性和效率,对辅助手部重建的物体特征提取同样具有显著效果。2)提出手部特征动态调整模块(dynamic adjustment module,DAM)。该模块引入外部注意力机制,对提取后的手部特征图先进行动态注意力调整,再输入进注意力模块,避免因输入噪声或复杂上下文关系导致权重分配不准确造成的干扰,得到更加精准的注意力全局权重。3)设计双流注意力协同特征增强模块(hand feature enhancement,HFE),首先将手部姿态特征信息和物体姿态特征信息进行融合,融合后再与经DAM调整后的特征进行再次融合,加权修正注意力中可能的错误聚焦,最终生成融合物体信息、细节、更丰富且更准确的手部优化特征图。

1 方 法

图1给出了本文总体网络结构,以手势姿态图像作为输入,通过针对性的手部特征提取模块(DS-FPN)进行双流特征提取,同时利用自顶向下和自底向上进行特征融合,使信息可以多次交互,提升特征表达能力。提取手部特征信息后,将手部特征的权重信息$\boldsymbol{Q}$注入注意力动态调整模块,引入外部注意力进行优化,得到优化的全局手部特征$\boldsymbol{A}_h$。此时,将手部特征的权重信息$\boldsymbol{Q}$、全局手部特征权重信息$\boldsymbol{A}_h$和物体特征权重信息$\boldsymbol{K}$、$\boldsymbol{V}$注入到双流注意力协同特征增强模块(HFE)进行注意力交互,最终获得更为准确的手部优化特征图。

1.1 双流手部特征金字塔网络

传统特征金字塔网络(FPN)通过单向自顶向下路径融合多尺度特征,虽然可以增强低层语义信息,但其单向传递机制导致手部细节(如指尖、关节纹理)在多次上采样中严重丢失,且缺乏对复杂遮挡场景的抗干扰能力。本文参照传统双向特征金字塔网

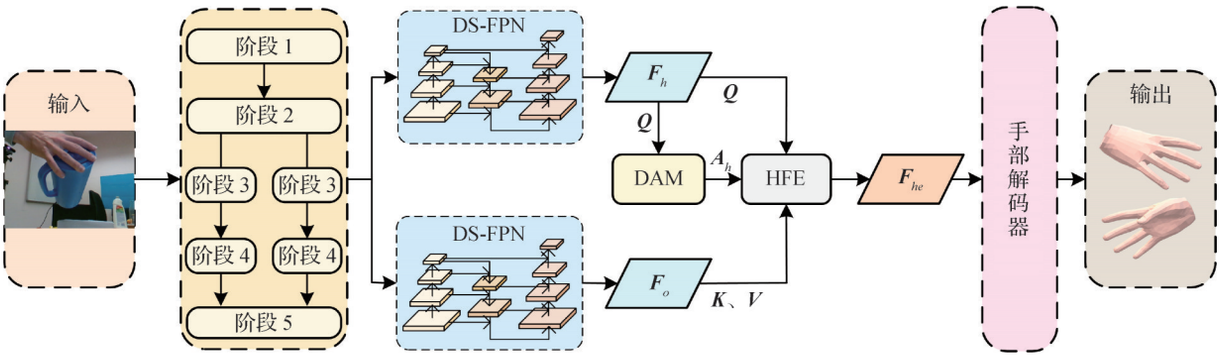

图1　总体网络架构

Fig. 1　Overall network architecture

络(bidirectional feature pyramid network, BiFPN),通过自顶向下与自底向上的双向路径融合多尺度特征,进一步提出双流手部特征金字塔网络(DS-FPN)。相较于BiFPN的单一融合路径,DS-FPN实现了特征的两次跨层交互。首先,自底向上流重点提取并强化局部细节特征;其次,自顶向下流有效注入高层语义和全局姿态信息;最后,第2次自底向上流对前两步获得的细节与全局信息进行深度融合,形成"细节→全局→细节+全局"的优化路径。本文设计的"自底向上→自顶而下→再自底向上"双路径结构与通道—空间双流交互机制,实现了手部特征的迭代优化与动态增强。

为克服传统BiFPN特征融合中空间信息建模不足的问题,本文给出一种增强型加权融合单元(mapping block),如图2所示,该单元由3部分组成,包括动态权重融合、卷积融合和非线性激活。

首先,对不同层级的输入特征图进行加权求和,权重可学习,用于自适应调整特征重要性,动态权重融合表达式为

$$\boldsymbol{F}_{\text{weight}}=\frac{\omega_1}{\omega_1+\omega_2+\varepsilon}\cdot\boldsymbol{F}_1+\frac{\omega_2}{\omega_1+\omega_2+\varepsilon}\cdot\boldsymbol{F}_2 \tag{1}$$

式中,ω_1和ω_2为可学习参数,初始化为1,通过反向传播优化;$\varepsilon=0.0001$,$\boldsymbol{F}_1$为来自高层特征的上采样结果,$\boldsymbol{F}_2$为来自当前层的横向连接特征。

其次,对加权融合后的特征进行空间增强与通道调整,通过1×1→3×3→1×1卷积链增强空间表示能力,表达式为

$$\boldsymbol{F}_{\text{fused}}=Conv_{1\times1}\left(Conv_{3\times3}\left(Conv_{1\times1}\left(\boldsymbol{F}_{\text{weight}}\right)\right)\right) \tag{2}$$

式中,1×1卷积降维用于压缩输入特征通道数,减少计算冗余。根据矩阵低秩理论,高维特征矩阵通常存在信息冗余,降维操作可近似保留主要特征模式,因此将加权融合后的特征通道数压缩至1/4,通过在低维空间进行高效计算,避免直接处理高维特征带来的参数量爆炸,显著减少计算冗余。3×3卷

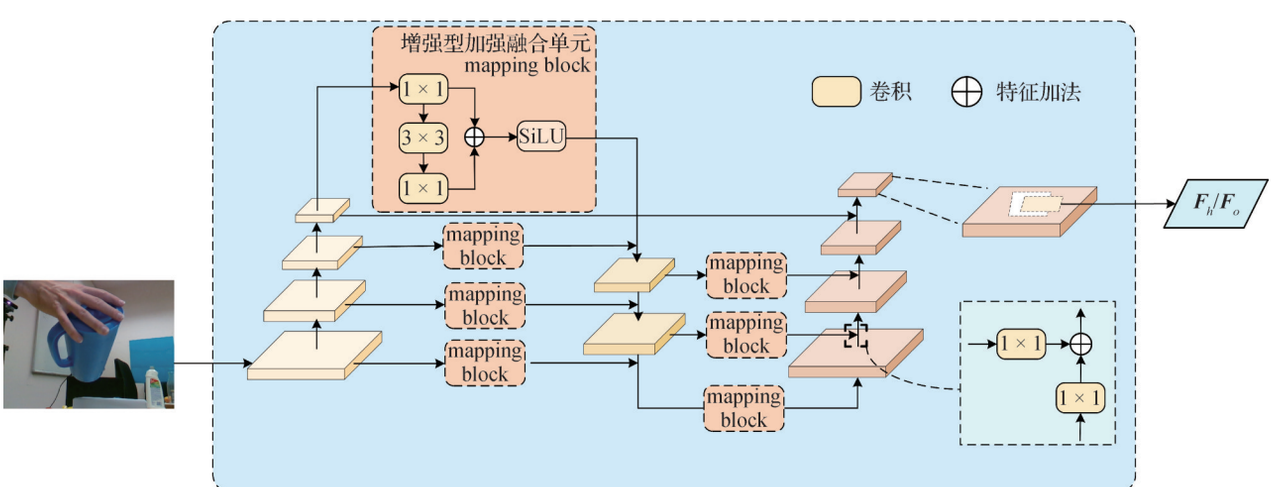

图2　双流手部特征金字塔网络

Fig. 2　Dual-stream hand-feature pyramid network

积进行空间增强是通过提取手部关节、指尖等局部空间特征完成，在降维后的低维空间中，通过 3 × 3 卷积核提取手部关节弯曲、指尖方向等局部空间特征。3 × 3 卷积核的局部感受野能够捕捉相邻像素间的空间相关性，1 × 1 卷积升维恢复则将特征通道数调整至目标维度，保持多尺度兼容性。该设计受到 Transformer 中多层感知机（multilayer perceptron，MLP）的启发，通过线性投影在低维与高维空间之间建立非线性映射，增强特征表达能力。将通道数恢复至目标维度，确保多尺度特征的兼容性。升维操作通过线性组合低维特征，重建高维表示，为后续层级提供适配不同分辨率任务的特征图。

最后，在卷积链末端引入 SiLU（sigmoid linear unit）激活函数，平衡梯度稳定性与非线性表达能力，其数学形式为

$$SiLU(x) = x \cdot \sigma(x) \tag{3}$$

式中，σ 为 sigmoid 函数。相较于传统 ReLU（rectified linear unit），sigmoid 项的引入使梯度在负区间非零，缓解了 ReLU 的“神经元死亡”问题，提升了模型训练的稳定性。通过 sigmoid 函数对特征图进行像素加权，增强关键区域，如指尖的响应强度，抑制低置信度区域。

1.2　手部特征动态调整模块

手部特征动态调整模块（DAM）进入特征增强前，先对手部特征提取的信息进行调整，并输入下一阶段的特征增强，为手部特征增强提供更多交互补充信息。手部特征动态调整模块基于外部注意力机制，对手部提取特征进行加工，输入手部特征图 $\boldsymbol{F}_h \in \mathbf{R}^{B\times C\times H\times W}$，经卷积层调整通道数并增强局部细节，输出全局手部特征 $\boldsymbol{A}_h \in \mathbf{R}^{B\times C_h\times H\times W}$，通过全局共享的记忆矩阵捕捉跨样本的共性特征，捕获手指边缘、关节位点等细粒度特征，如图 3 所示。

首先，输入手部特征图 $\boldsymbol{F}_h \in \mathbf{R}^{B\times C\times H\times W}$，通过卷积层提取局部细节，并调整通道维度以适配后续全局注意力模块，具体表达式为

$$\boldsymbol{Q}_s = ReLU\left(\boldsymbol{W}_{\text{conv}} * \boldsymbol{Q} + \boldsymbol{b}_{\text{conv}}\right) \tag{4}$$

式中，$\boldsymbol{W}_{\text{conv}}$ 为 3 × 3 卷积核权重，* 代表卷积运算，$\boldsymbol{b}_{\text{conv}}$ 为偏置项，ReLU 激活函数增强非线性表达能力，输出特征 $\boldsymbol{Q}_s \in \mathbf{R}^{B\times C_h\times H\times W}$ 将作为后续全局注意力模块的输入。该层通过局部感受野提取细粒度特征，为跨样本依赖建模提供基础。

为建模手部特征的全局依赖关系并减少计算开销，本文提出一种基于外部记忆的跨样本全局注意力机制，该层通过可学习的键—值记忆矩阵（$\boldsymbol{M}_K$、$\boldsymbol{M}_V$）动态聚合跨样本的共性特征，其计算过程定义如下：

1）定义两组全局共享的可学习参数。键记忆矩阵 $\boldsymbol{M}_K \in \mathbf{R}^{M\times C_1}$，编码数据集中手部姿态的通用键特征，每个记忆单元对应一种潜在的手部结构先验。值记忆矩阵 $\boldsymbol{M}_V \in \mathbf{R}^{M\times C_2}$ 及存储与键特征对应的语义响应用于动态调整特征重要性。其中，M 为记忆单元数量（$M \ll H \times W$），C_1、C_2 分别为输入与输出通道数。

2）与 $\boldsymbol{M}_K$ 交互，计算注意力权重。将局部特征提取层输出的特征图 $\boldsymbol{Q}_s \in \mathbf{R}^{B\times C_1\times H\times W}$ 展开为 $\mathbf{R}^{B\times C_1\times N}$（$N = H \times W$），计算其与 $\boldsymbol{M}_K$ 的相似性，计算注意力权重的表达式为

$$\boldsymbol{A}_{\text{raw}} = \boldsymbol{Q}_S^{\mathrm{T}} \times \boldsymbol{M}_k \in \mathbf{R}^{B\times N\times W} \tag{5}$$

3）归一化。使用 LayerNorm 调整维度顺序，生成归一化注意力权重 $\boldsymbol{A} \in \mathbf{R}^{B\times N\times M}$

$$\boldsymbol{A} = LayerNorm\left(\boldsymbol{A}_{\text{raw}}\right) \tag{6}$$

4）与 $\boldsymbol{M}_V$ 交互，特征动态聚合。将注意力权重 $\boldsymbol{A}$ 与值记忆矩阵 $\boldsymbol{M}_V$ 加权求和，生成全局增强后的特征图 $\boldsymbol{A}_h \in \mathbf{R}^{B\times C_2\times H\times W}$，具体为

$$\boldsymbol{A}_h = \boldsymbol{A} \cdot \boldsymbol{M}_v \tag{7}$$

该过程通过外部记忆矩阵隐式编码跨样本的手部姿态共性（如关节弯曲模式），同时以线性复杂度实现高效计算。

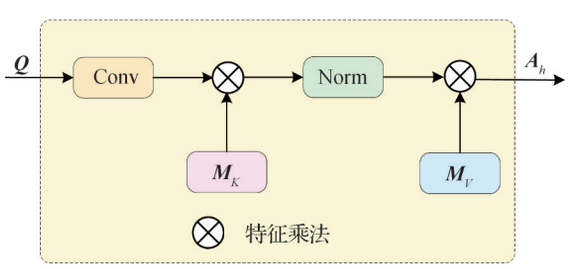

图 3　手部特征动态调整模块

Fig. 3　Hand feature dynamic adjustment module

1.3　双流注意力协同特征增强模块

在复杂交互场景（如手部抓握物体、遮挡干扰）中，传统手部特征提取方法往往难以兼顾局部细节刻画与全局语义关联。

针对这一挑战，本文提出手部特征增强模块（HFE），如图 4 所示，通过跨模态注意力引导与多粒

度特征融合的协同设计，实现高效特征优化。该方法以动态调整模块输出的全局手部特征$\boldsymbol{A}_h$为基础，引入物体特征作为跨模态上下文引导源，通过多头注意力机制建立手部—物体空间关联，使特征响应自适应聚焦于交互关键区域（如工具抓握点、遮挡边界）。

为进一步平衡特征表征的鲁棒性与判别性，模块采用双路径融合策略：一方面保留原始特征细节以维持底层信息完整性；另一方面通过层归一化与多层感知机细化高阶语义特征，抑制背景噪声干扰。

首先，对手部特征图$\boldsymbol{F}_h$和物体特征图$\boldsymbol{F}_o$进行特征编码与序列化处理，通过1×1卷积调整通道数，生成查询矩阵$\boldsymbol{Q}\in\mathbf{R}^{B\times C\times H\times W}$、键矩阵$\boldsymbol{K}\in\mathbf{R}^{B\times C\times H_0\times W_0}$和值矩阵$\boldsymbol{V}\in\mathbf{R}^{B\times C\times H_0\times W_0}$。具体为

$$\boldsymbol{Q}=Conv_{1\times1}\left(\boldsymbol{F}_h\right) \tag{8}$$

$$\boldsymbol{K}=Conv_{1\times1}\left(\boldsymbol{F}_o\right) \tag{9}$$

$$\boldsymbol{V}=Conv_{1\times1}\left(\boldsymbol{F}_o\right) \tag{10}$$

其次，基于查询矩阵$\boldsymbol{Q}$、键矩阵$\boldsymbol{K}$和值矩阵$\boldsymbol{V}$进行多头注意力计算，其中，第i个单头注意力$\boldsymbol{head}_i$的计算具体为

$$\boldsymbol{head}_i=softmax\left(\frac{\boldsymbol{q}_i\boldsymbol{k}_i^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{v}_i,\ d_k=\frac{C}{h} \tag{11}$$

式中，$\boldsymbol{q}_i$为将查询矩阵$\boldsymbol{Q}$划分为h个组得到的第i个查询子矩阵，$\boldsymbol{k}_i$为将键矩阵$\boldsymbol{K}$划分为h个组得到的第i个键子矩阵，$\boldsymbol{v}_i$为将值矩阵$\boldsymbol{V}$划分为h个组得到的第i个值子矩阵，C为特征图$\boldsymbol{Q}$的通道数，d_k为每个组的通道维度。

对h个单头注意力进行拼接，经过线性投影后与全局手部特征$\boldsymbol{A}_h$按元素相加，得到手物协同特征$\boldsymbol{F}_f$。具体为

$$\boldsymbol{A}_{\text{attn}}=Concat\left(\boldsymbol{head}_1,\cdots,\boldsymbol{head}_h\right)\boldsymbol{W}_0 \tag{12}$$

$$\boldsymbol{F}_f=\boldsymbol{A}_h+\boldsymbol{A}_{\text{attn}} \tag{13}$$

将手物协同特征作为输入，与动态提取特征融合，进行双分支细化与特征增强。其中，分支A为残差保留，用于保留初步融合特征和维持信息完整性。具体为

$$\boldsymbol{F}_a=\boldsymbol{F}_f \tag{14}$$

分支B经过MLP初步细化后，进入层归一化操作，稳定训练过程，缓解梯度爆炸，之后进入多层感知机，通过非线性变换细化特征。

$$\boldsymbol{F}_{\text{norm}}=LayerNorm\left(\boldsymbol{F}_f\right) \tag{15}$$

$$\boldsymbol{F}_{\text{mlp}}=MLP\left(\boldsymbol{F}_{\text{norm}}\right) \tag{16}$$

式中，MLP由两个线性层与GELU（Gaussian error linear unit）激活函数构成。具体为

$$MLP\left(\boldsymbol{F}_{\text{norm}}\right)=\boldsymbol{W}_2\cdot GELU\left(\boldsymbol{W}_1\cdot\boldsymbol{F}_{\text{norm}}+\boldsymbol{b}_1\right)+\boldsymbol{b}_2 \tag{17}$$

式中，$\boldsymbol{W}_1$为特征扩展，通过将维度从C扩展至$4C$，增强模型容量，捕捉复杂的非线性关系，如手部关节间的动态关联。$\boldsymbol{W}_2$为特征压缩，将维度恢复至C，防止参数量爆炸，同时过滤冗余信息。$\boldsymbol{b}_1$和$\boldsymbol{b}_2$均为偏置项。

最终，进行融合，将A和B两分支结果相加，生成优化后的手部特征图。即

$$\boldsymbol{F}_{\text{opt}}=\boldsymbol{F}_a+\boldsymbol{F}_{\text{mlp}} \tag{18}$$

手部特征增强模块（HFE）采用两阶段自适应融合。阶段1通过跨模态注意力将物体语义（如工具形状、抓握方向）注入手部特征，增强交互区域（如指尖、掌心）的响应强度。在阶段2中，双分支结构平衡特征保留与细化，分支A防止信息丢失，分支B抑制噪声并增强判别性。最终，获得融合了物体信息的手部细节更丰富、更准确的手部优化特征图。

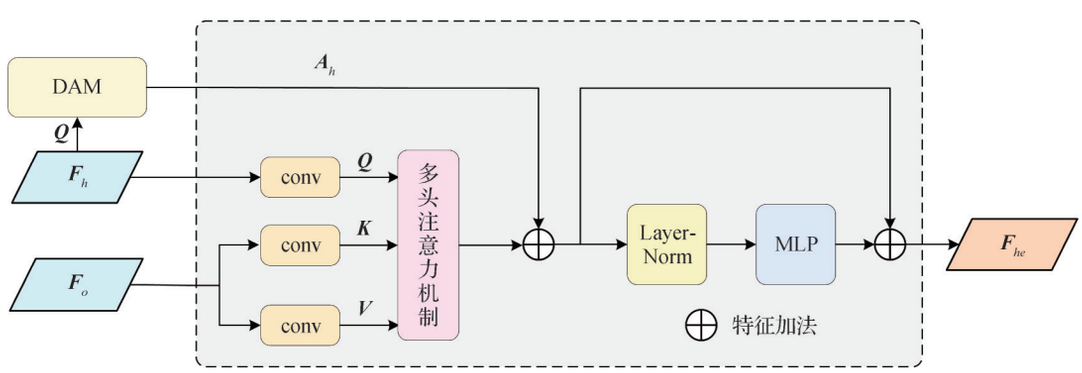

图4　双流注意力协同特征增强模块

Fig. 4　Double-stream attention coordination feature enhancement module

1.4 损失函数

针对手部姿态估计任务，本文提出基于物体协同监督的损失函数设计。尽管核心目标仍为手部姿态估计，但引入物体姿态估计作为协同约束，通过跨模态特征交互增强手部关节的空间定位能力。因此，损失函数仍需设计物体协同损失，总损失函数由手部基础损失与物体协同损失构成，即

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{hand}} + \varphi\mathcal{L}_{\text{obj}} \tag{19}$$

式中，$\mathcal{L}_{\text{hand}}$ 为手部姿态主监督项（涵盖2D/3D坐标及MANO参数约束）。$\mathcal{L}_{\text{obj}}$ 为物体姿态协同监督项（如物体包围盒投影误差）。$\varphi(\varphi < 1)$为协同权重系数，控制物体信息对手部优化的影响强度。

损失函数由手部姿态损失与物体姿态损失两部分构成：前者通过计算关节角度偏差与空间位置误差直接优化手部姿态精度，后者基于手—物接触区域的几何一致性建模交互合理性。二者通过加权融合形成联合优化目标，在强化手部特征学习的同时，利用物体姿态信息约束手部运动的物理可行性，从而有效提升遮挡场景下姿态估计的鲁棒性与空间一致性。

手部姿态损失由3部分组成，包括2D关节点检测损失 $\mathcal{L}_{\text{H}}$、3D顶点与关节点损失 $\mathcal{L}_{3\text{d}}$ 和 MANO（model of articulated neutral object）参数损失 $\mathcal{L}_{\text{mano}}$，具体为

$$\mathcal{L}_{\text{hand}} = \alpha_{\text{h}}\mathcal{L}_{\text{H}} + \alpha_{3\text{d}}\mathcal{L}_{3\text{d}} + \alpha_{\text{mano}}\mathcal{L}_{\text{mano}} \tag{20}$$

式中，α_{h}，$\alpha_{3\text{d}}$，α_{mano} 为平衡每个损失函数权重的系数。

$\mathcal{L}_{\text{H}}$ 为2D关节点损失，用于监督手部2D关节点的坐标预测，采用均方误差（mean square error，MSE），对每个Batch中的每个关节点计算预测坐标与真实坐标的L2损失（均方误差），并取平均。具体为

$$\mathcal{L}_{\text{H}} = \frac{1}{B \cdot N_h}\sum_{b=1}^{B}\sum_{i=1}^{N_h}\left\| \hat{\boldsymbol{p}}_i^{(b)} - \boldsymbol{p}_i^{(b)} \right\|_2^2 \tag{21}$$

式中，B 为batchsize大小，$N_h = 21$ 为手部关节点数量，$\hat{\boldsymbol{p}}_i^{(b)}$ 为第 b 个样本中第 i 个关节点的预测2D坐标，$\boldsymbol{p}_i^{(b)}$ 为第 b 个样本中第 i 个关节点的真实2D坐标。

3D顶点与关节点损失是顶点L2损失与关节点L2损失按各自数量归一化后的加权和，其中顶点L2损失计算所有3D顶点预测值与真实值的差异，关节点L2损失计算所有3D关节点预测值与真实值的差异。具体为

$$\mathcal{L}_{3\text{d}} = \frac{1}{B \cdot V}\sum_{b=1}^{B}\sum_{j=1}^{V}\left\| \hat{\boldsymbol{v}}_j^{(b)} - \boldsymbol{v}_j^{(b)} \right\|_2^2 + \frac{1}{B \cdot N_h}\sum_{b=1}^{B}\sum_{i=1}^{N_h}\left\| \hat{\boldsymbol{j}}_i^{(b)} - \boldsymbol{j}_i^{(b)} \right\|_2^2 \tag{22}$$

式中，$\boldsymbol{V} = 778$ 为MANO模型定义的手部3D顶点数量，$\hat{\boldsymbol{v}}_i^{(b)}$ 为第 b 个样本中第 i 个顶点的预测3D坐标，$\boldsymbol{v}_j^{(b)}$ 为第 b 个样本中第 j 个顶点的真实3D坐标，$\hat{\boldsymbol{j}}_i^{(b)}$ 为第 b 个样本中第 i 个关节点的预测3D坐标，$\boldsymbol{j}_i^{(b)}$ 为第 b 个样本中第 j 个关节点的真实3D坐标。$\hat{\boldsymbol{\beta}}^{(b)}$ 为第 b 个样本中预测的形状参数，$\boldsymbol{\beta}^{(b)}$ 为第 b 个样本中真实的形状参数，$\hat{\boldsymbol{\theta}}^{(b)}$ 为第 b 个样本中预测的姿势参数，$\boldsymbol{\theta}^{(b)}$ 为第 b 个样本中真实的姿势参数。

MANO参数损失分别计算形状参数 $\boldsymbol{\beta}$ 与姿态参数 $\boldsymbol{\theta}$ 的L2损失，按Batch取平均，具体为

$$\mathcal{L}_{\text{mano}} = \frac{1}{B}\sum_{b=1}^{B}\left(\left\| \hat{\boldsymbol{\beta}}^{(b)} - \boldsymbol{\beta}^{(b)} \right\|_2^2 + \left\| \hat{\boldsymbol{\theta}}^{(b)} - \boldsymbol{\theta}^{(b)} \right\|_2^2 \right) \tag{23}$$

物体姿态协同损失包括2D投影损失与置信度监督损失，具体为

$$\mathcal{L}_{\text{obj}} = \alpha_{\text{p2d}}\mathcal{L}_{\text{p2d}} + \alpha_{\text{conf}}\mathcal{L}_{\text{conf}} \tag{24}$$

式中，α_{p2d} 和 α_{conf} 是平衡每个损失函数权重的系数。$\mathcal{L}_{\text{p2d}}$ 为2D投影损失，约束物体关键点的2D投影位置，采用平均绝对误差（mean absolute error，MAE）。具体为

$$\mathcal{L}_{\text{p2d}} = \frac{1}{B \cdot N_0}\sum_{b=1}^{B}\sum_{k=1}^{N_o}\left\| \hat{\boldsymbol{o}}_k^{(b)} - \boldsymbol{o}_k^{(b)} \right\|_1 \tag{25}$$

式中，$N_0 = 8$ 为物体3D包围盒的角点数量。$\hat{\boldsymbol{o}}_k^{(b)}$ 为第 k 个角点的预测2D投影坐标，$\boldsymbol{o}_k^{(b)}$ 为对应的真实投影坐标。

$\mathcal{L}_{\text{conf}}$ 是置信度损失，监督物体的存在置信度（0表示不可见，1表示可见），具体为

$$\mathcal{L}_{\text{conf}} = \frac{1}{B}\sum_{b=1}^{B}\left\| \hat{\boldsymbol{c}}^{(b)} - \boldsymbol{c}^{(b)} \right\|_1 \tag{26}$$

式中，$\hat{\boldsymbol{c}}^{(b)}$ 为预测的物体存在置信度，范围为(0,1)；$\boldsymbol{c}^{(b)}$ 为真实二值标签，范围为(0,1)。

2 实 验

2.1 实验数据集与评价指标

实验采用 Dex-YCB（dexterous-YCB）和 HO3D（hand-object 3D）两个公开数据集进行手部姿态估计的验证。Dex-YCB数据集上专注于手部精细姿态估计与动态手—物交互研究。数据集包含来自10类YCB物体的动态抓取序列，涵盖抓握、平移、旋转及释放等完整交互动作，训练集约500 000帧，测试集包含5 000帧，覆盖多样化的操作者与物体实例组合。所有数据通过高精度光学运动捕捉系统（如

Vicon)与多台同步RGB-D摄像头(如Intel RealSense)采集,确保手部与物体位姿的ms级精度,同时提供多视角图像与深度信息以增强三维重建的鲁棒性。标注信息基于MANO参数化手部模型生成,包含45维姿态参数与10维形状参数,物体6D位姿则通过刚体配准技术精确标注,支持手—物协同位姿的联合分析。官方采用S0—S2分层划分策略:S0按物体实例划分,以验证模型对未见物体的泛化性;S1按操作者划分,以评估跨用户适应性;S2混合划分,用于综合性能评测。研究者需通过指定平台提交结果以确保评估一致性。

该数据集提供MPJPE(mean per joint position error)与PA-MPJPE(procrustes-aligned mean per joint position error)双指标,其中MPJPE直接计算未对齐的全局关节位置误差以反映绝对空间精度,而PA-MPJPE通过Procrustes对齐消除全局位姿偏移,专注于评估局部手部姿态的几何一致性,两者结合可全面衡量模型在真实场景中的实用性能。

HO3D数据集专注于复杂手—物交互3D感知基准数据集,旨在推动动态场景下手部姿态与物体位姿的联合估计研究。其包含多视角同步采集的RGB-D图像序列(尺寸为640×480像素),提供21个关节手部3D姿态标注(基于MANO模型)和交互物体的6D位姿(3D平移+旋转),覆盖12类常见物体(如剪刀、杯子、扳手)的多样化抓取动作。数据集划分为66 000帧训练集与11 000帧测试集,模拟真实场景中的重度遮挡(手—物遮挡率超70%)、快速手部运动(指尖速度>20 cm/s)及光照突变等挑战。

对于手部姿态估计,评价指标分为两类:

1)手部姿态误差。采用MPJPE(平均关节位置误差,毫米级)与PA-MPJPE(刚性对齐后误差),衡量3D关节点定位精度。

2)重建精度:采用手部姿态估计领域的标准指标F@5与F@15,即预测的手部关节点与真实关节点之间的欧氏距离在5 mm与15 mm阈值内的比例,用以衡量手部自身几何形状的准确性。

测试集标注严格保密,需通过官方在线评估服务器提交预测结果,确保评测过程公平可靠,该机制使其成为手—物交互领域最具权威性的技术验证平台之一。

2.2 实验环境

本实验采用NVIDIA A10 Tensor Core显卡,操作系统选择Windows操作系统。开发框架选用PyTorch。在训练过程中,采用Adam优化器和5×10^{-4}的权重衰减进行优化。批量大小(batch-size)设置为64,训练的总次数是60。初始学习率为10^{-4},每10个epoch衰减一次。

2.3 实验结果与分析

2.3.1 Dex-YCB数据集指标评价结果及分析

在Dex-YCB数据集的实验结果如表1所示,对比模型结果均为官方论文所给数据结果。HOCEN(本文)模型在MPJPE和PA-MPJPE两项关键指标上均超越了多个先进的网络模型,展现出显著的优势,取得了更好的效果。

具体而言,HOCEN模型在MPJPE指标上比SemGCN(semantic graph convolutional network)(Shuang等,2024)精度提升0.8 mm,比HOFEC(顾思远和高曙,2025)精度提升0.3 mm,在降低误差方面展现出显著优势。同时,在PA-MPJPE指标上,HOCEN模型比SemGCN(Shuang等,2024)精度提升0.2 mm,并比SimpleHand(Zhou等,2024)和HOFEC(顾思远和高曙,2025)精度提升0.1 mm,实现了最低误差,进一步验证了模型在对手势姿态几何对齐及细节拟合方面的卓越能力。

从整体对比来看,HOCEN已在Dex-YCB数据集上取得了超越SimpleHand(Zhou等,2024)等模型的最优效果,证明了其在这一任务中的优秀竞争力。这一结果不仅展示了模型在复杂场景(如遮挡、光照变化、手势复杂变形等)下的强泛化能力与鲁棒性,也反映了模型在网络结构优化、特征提取策略以及基于几何约束的训练方法等方面的创新之处。

综上所述,HOCEN模型在多个关键指标上均达到了当前领先水平,为手势姿态估计领域提供了新的思路和实践依据。模型在提升整体精度的同时,兼顾了低误差与细节捕捉,充分证明了其在实际应用中的潜力与广泛适用性。

2.3.2 HO3D数据集指标评价结果及分析

为验证模型在复杂手—物交互场景下的鲁棒性,本节对比所提方法与当前主流算法在HO3D测试集上的性能,对比结果如表2所示。

在HO3D数据集上的对比实验中,本文提出的HOCEN模型在多项指标上均展现出优异性能,与其他先进方法相比具有明显优势。具体来看,HOCEN模型在关节点误差(Joint)和网格误差(Mesh)上分别

表 1　Dex-YCB 数据集上 HOCEN（本文）与其他模型的手势姿态估计结果对比

Table 1　Comparison of gesture pose estimation results between HOCEN and other models on the Dex-YCB dataset

模型	MPJPE/mm ↓	PA-MPJPE/mm ↓
Spurr 等人（2020）	17.3	6.8
Liu 等人（2021）	15.3	6.6
HandOccNet（Park 等，2022）	14.0	5.8
HFL（Lin 等，2023）	12.6	5.5
SemGCN（Shuang 等，2024）	13.2	5.6
SimpleHand（Zhou 等，2024）	**12.4**	5.5
HOFEC（顾思远和高曙，2025）	12.7	5.5
HOCEN（本文）	**12.4**	**5.4**

注：加粗、下划线字体表示各列最优、次优结果。“↓”表示值越低越好。

表 2　HO3D 数据集上 HOCEN（本文）与其他模型的手势姿态估计结果对比

Table 2　Comparison of gesture pose estimation results between HOCEN (in this paper) and other models on the HO3D dataset

模型	Joint ↓	Mesh ↓	F@5/% ↑	F@15/% ↑
Choi 等人（2020）	12.5	12.7	44.1	90.9
Hasson 等人（2022）	11.4	11.4	42.8	93.2
Moon 和 Lee(2020)	11.2	13.9	40.9	93.2
Hampali 等人（2020）	10.7	10.6	50.6	94.2
Lin 等人（2021）	10.4	11.1	48.4	94.6
Liu 等人（2021）	10.1	9.7	53.2	95.2
Yang 等人（2022）	11.4	10.9	48.8	94..4
Chen 等人（2022）	**9.2**	9.4	53.8	59.7
Lin 等人（2023）	9.4*	9.3*	54.6*	95.9*
HOCEN（本文）	**9.2**	**9.1**	**56.1**	**96.1**

注：加粗、下划线字体表示各列最优、次优结果。“*”表示在本文实验环境下复现的结果。“↑”和“↓”分别表示值越高、越低越好。

达到 9.2 和 9.1，实现了极低的误差，与表现出色的 Chen 等人（2022）模型持平或更优。同时，在 F@15 指标上，HOCEN 模型以 96.1% 的得分略微超越了 Lin 等人（2023）模型的 95.9%，表现出在高容错阈值下具有更强的稳定性。此外，HOCEN 模型在 F@5 指标上取得 56.1%，比 Lin 等人（2023）的模型仍有进一步的提升，且显著优于更早期的方法（如 Hampali 和 Hasson）。整体来看，HOCEN 模型在降低误差的同时，兼顾了高精度和高稳定性，展现出在手势姿态估计任务中的强大泛化能力和整体优势。

2.3.3　可视化结果对比展示

为了更直观地展示 HFE 模型在手势姿态估计任务中的优越性能，本文对模型的预测结果进行了可视化分析。通过将预测手势与真实手势进行对比，可以清晰地观察到 HFE 模型在关节点定位、网格重建以及整体姿态拟合方面的准确性与细节表现。此外，为了进一步突出模型在复杂场景下的鲁棒性，选取了不同挑战场景（如遮挡、光照变化等）进行可视化展示，从而验证模型在真实应用环境中的泛化能力和稳定性。

图 5 和图 6 分别为 Dex-YCB 数据集和 HO3D 数据集的可视化对比图。

2.4　消融实验

为验证本文提出方法的有效性，在 Dex-YCB 与

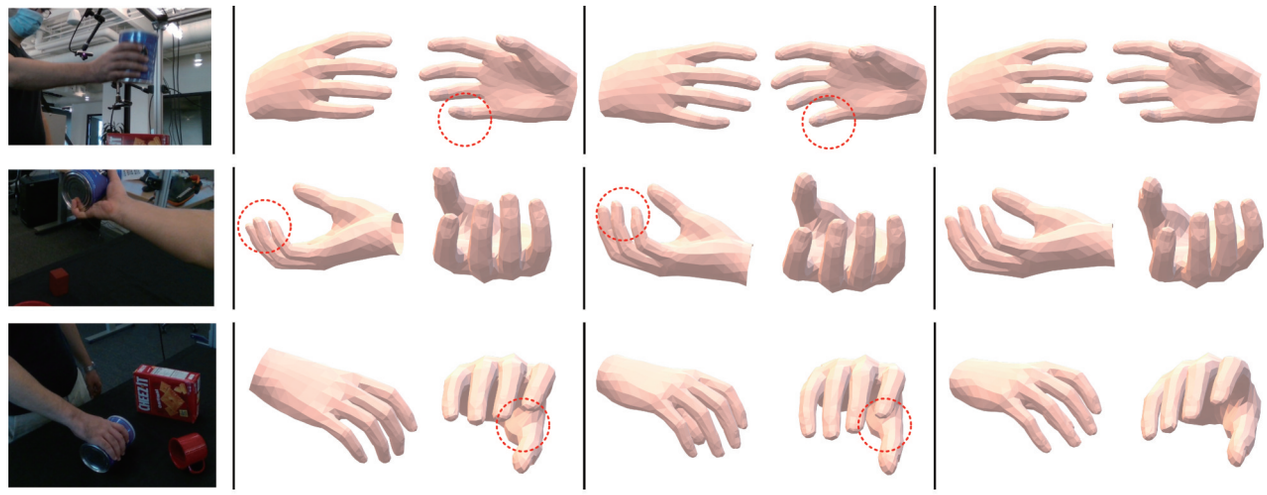

图5　Dex-YCB数据集可视化

Fig. 5　Visualization of Dex-YCB dataset ((a) input(ours); (b) HOCEN; (c) HFL-Net; (d) HandOccNet)

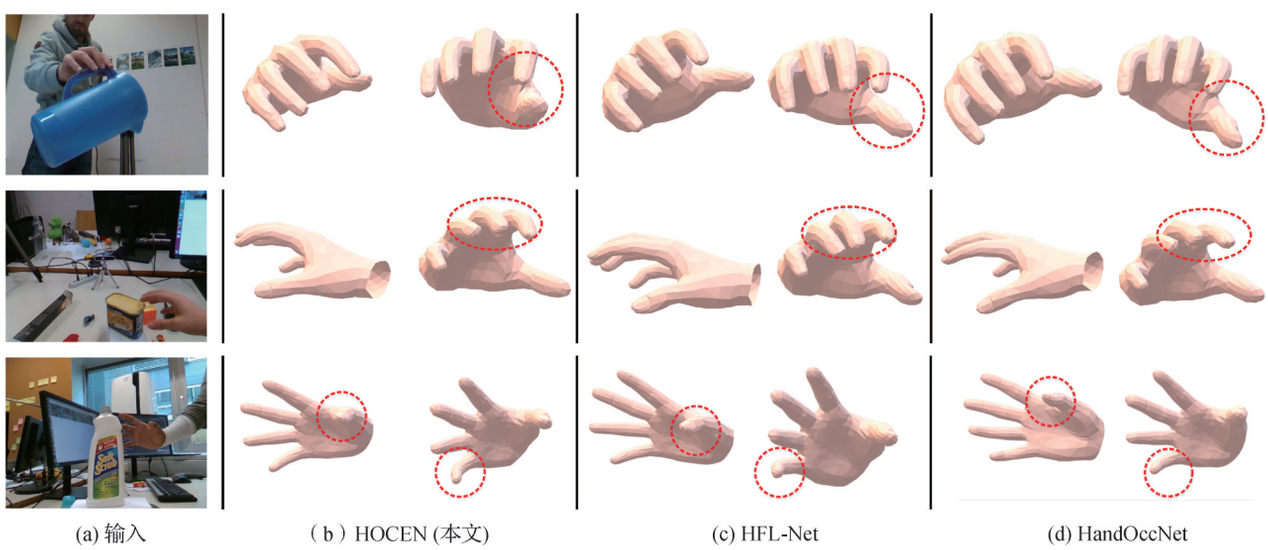

图6　HO3D数据集可视化

Fig. 6　Visualization of HO3D dataset ((a) input; (b) HOCEN(ours); (c) HFL-Net; (d) HandOccNet)

HO3D标准数据集上展开系统性验证。采用渐进式模块集成策略，以传统特征金字塔网络(FPN)为基线模型，依次引入双流手部特征金字塔网络(DS-FPN)、手部特征动态调整(DAM)和双流注意力协同特征增强(HFE) 3个核心模块。实验结果如表3和表4所示。实验结果表明，DS-FPN模块通过跨尺度注意力机制使平均关节位置误差(MPJPE)从基线12.56 mm降低至12.47 mm，DAM模块在遮挡场景下进一步将MPJPE优化至12.44 mm，HFE模块最终实现5.45 mm至5.42 mm的精度提升，特别是在拇指根部等高遮挡区域误差显著降低。实验证实，DS-FPN、DAM与HFE的模块设计在多尺度特征建模、动态交互适应与跨模态语义对齐方面展现出互补优势，使模型在精度与鲁棒性层面均获得实质性提升。

为更直观对比各模块依次叠加后在手势姿态估计任务中的性能提升，本文分别对Dex-YCB与HO3D数据集上的结果绘制了对应数据集相应指标的曲线图，如图7—图9所示。可以看出，逐层叠加的模块在Dex-YCB与HO3D数据集上的指标均有明显的提升，各模块协同工作，使得手势估计在精度、鲁棒性和效率均获得显著提升。

表 3　Dex-YCB 数据集上 HOCEN（本文）与模块叠加方法的手势姿态估计对比

Table 3　Comparison of gesture pose estimation between HOCEN (this paper) and module superposition method on the Dex-YCB dataset

模型	MPJPE ↓	PA-MPJPE ↓
+FPN	12.56	5.47
+DS-FPN	12.47	5.47
+DS-FPN+DAM	12.44	5.45
+DS-FPN+DAM+HFE	12.44	5.42

注：“↓”表示值越低越好。

表 4　HO3D 数据集上 HOCEN（本文）与模块叠加方法的手势姿态估计对比

Table 4　Comparison of gesture pose estimation between HOCEN (this paper) and module superposition method on the HO3D dataset

模型	Joint ↓	Mesh ↓	F@5 /% ↑	F@15 /% ↑
+FPN	9.4	9.3	54.6	95.9
+DS-FPN	9.4	9.2	55.2	95.9
+DS-FPN+DAM	9.3	9.1	55.8	96.0
+DS-FPN+DAM+HFE	9.2	9.1	56.1	96.1

注：“↑”和“↓”分别表示值越高、越低越好。

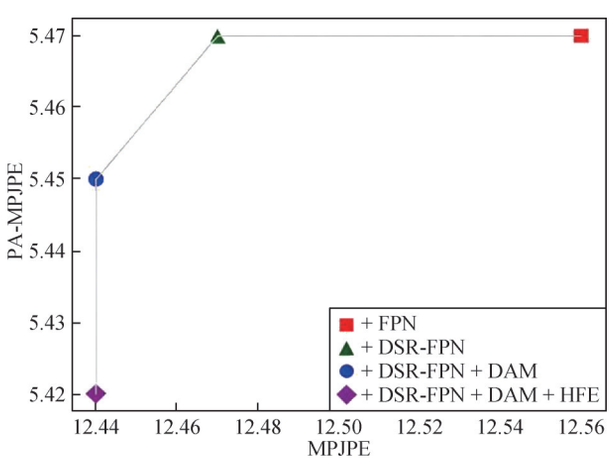

图 7　Dex-YCB 数据集 MPJPE 与 PA-MPJPE 指标对比

Fig. 7　Comparison of MPJPE and PA-MPJPE metrics on the Dex-YCB dataset

3　结　论

本文提出一种融合手物特征的三维手部姿态估计网络（HOCEN），通过多层次特征交互与跨模态注意力机制，显著提升了复杂交互场景下的手部姿态估计精度。针对现有方法在多尺度特征融合、遮挡鲁棒性及手—物特征干扰等方面的不足，本文方法创新性地设计了双流特征金字塔模块，通过动态尺度感知与跨层级信息聚合，有效缓解了传统金字塔结构的特征退化问题；引入基于外部记忆的动态调整机制，在遮挡区域自适应增强手部语义表达，抑制背景噪声干扰；进一步构建双流协同注意力模块，以双向特征引导策略实现手—物几何约束与语义互补的深度融合。实验表明，该方法在 HO3D 与 Dex-YCB 数据集上的综合性能优于主流方法，尤其在重度遮挡与复杂交互场景下展现出更强的稳定性与细节还原能力。消融实验验证了各模块对手部关键点

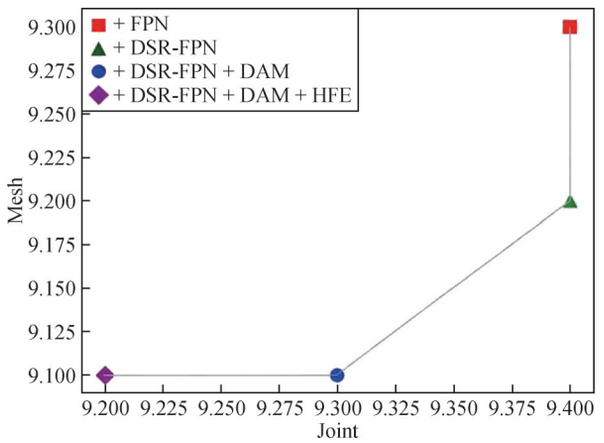

图 8　HO3D 数据集 Joint 与 Mesh 的指标对比

Fig. 8　Comparison of metrics between Joint and Mesh on the HO3D dataset

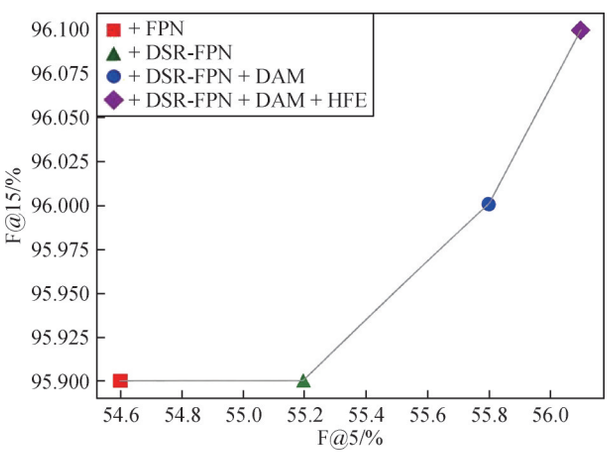

图 9　HO3D 数据集 F@5 与 F@15 的指标对比

Fig. 9　Comparison of metrics between F@5 and F@15 on the HO3D dataset

定位精度与物体空间关系建模的协同增益。然而，当前模型对多手多物体交互场景的适应性有限，且计算效率有待进一步优化。未来工作将探索轻量化架构设计与时序建模策略，以扩展方法在动态多对象交互任务中的实用性。

参考文献（References）

Chen X Y, Liu Y F, Dong Y J, Zhang X, Ma C Y, Xiong Y M, et al. 2022. MobRecon: mobile-friendly hand mesh reconstruction from monocular image//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 20512-20522 [DOI: 10.1109/CVPR52688.2022.01989]

Chen Y J, Tu Z G, Kang D, Chen R Z, Bao L C, Zhang Z Y, et al. 2021. Joint hand-object 3D reconstruction from a single image with cross-branch feature fusion. IEEE Transactions on Image Processing, 30: 4008-4021 [DOI: 10.1109/TIP.2021.3068645]

Chen Y L, Wang Z C, Peng Y X, Zhang Z Q, Yu G and Sun J. 2018. Cascaded pyramid network for multi-person pose estimation//Proceedings of 2018 CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7103-7112 [DOI: 10.1109/CVPR.2018.00742]

Choi H, Moon G and Lee K M. 2020. Pose2Mesh: graph convolutional network for 3D human pose and mesh recovery from a 2D human pose//Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK: Springer: 769-787 [DOI: 10.1007/978-3-030-58571-6_45]

Gu S Y and Gao S. 2025. Hand-object pose estimation method based on fusion feature enhancement and complementary. Journal of Image and Graphics, 30(5): 1433-1449 (顾思远, 高曙. 2025. 融合特征增强与互补的手物姿态估计方法. 中国图象图形学报, 30(5): 1433-1449) [DOI: 10.11834/jig.240272]

Hasson Y, Tekin B, Bogo F, Laptev I, Pollefeys M and Schmid C. 2020. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 571-580 [DOI: 10.1109/CVPR42600.2020.00065]

Hampali S, Rad M, Oberweger M and Lepetit V. 2020. HOnnotate: a method for 3D annotation of hand and object poses//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 3193-3203 [DOI: 10.1109/CVPR42600.2020.00326]

He Y S, Huang H B, Fan H Q, Chen Q F and Sun J. 2021. FFB6D: a full flow bidirectional fusion network for 6D pose estimation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 3002-3012 [DOI: 10.1109/CVPR46437.2021.00302]

Jia D, Li Y Y, An T and Zhao J Y. 2023. Complex gesture pose estimation network fusing multiscale features. Journal of Image and Graphics, 28(9): 2887-2898 (贾迪, 李宇扬, 安彤, 赵金源. 2023. 融合多尺度特征的复杂手势姿态估计网络. 中国图象图形学报, 28(9): 2887-2898) [DOI: 10.11834/jig.220636]

Lee K W, Liu S H, Chen H T and Ito K. 2019. Silhouette-net: 3d hand pose estimation from silhouettes [EB/OL]. [2025-07-01]. https://arxiv.org/pdf/2206.03001.pdf

Liu S W, Jiang H W, Xu J R, Liu S F and Wang X L. 2021. Semi-supervised 3D hand-object poses estimation with interactions in time//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 14682-14692 [DOI: 10.1109/CVPR46437.2021.01445]

Liu S, Wu W, Wu J and Lin Y. 2022. Spatial-temporal parallel transformer for arm-hand dynamic estimation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 20523-20532.

Lin K, Wang L J and Liu Z C. 2021. End-to-end human pose and mesh reconstruction with transformers//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 1954-1963 [DOI: 10.1109/CVPR46437.2021.00199]

Lin T Y, Dollár P, Girshick R, He K M, Hariharan B and Belongie S. 2017. Feature pyramid networks for object detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 936-944 [DOI: 10.1109/CVPR.2017.106]

Lin Z, Ding C, Yao H, Kuang Z and Huang S. 2023. Harmonious feature learning for interactive hand-object pose estimation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 12989-12998

Moon G and Lee K M. 2020. I2l-meshnet: image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image//Proceedings of the 16th European Conference on Computer Vision Glasgow.UK:Springer: 752-768

Park J, Oh Y, Moon G, Choi H and Lee K M. 2022. HandOccNet: occlusion-robust 3D hand mesh estimation network//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 1486-1495 [DOI: 10.1109/CVPR52688.2022.00155]

Shuang F, He W and Li S. 2024. 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

Spurr A, Iqbal U, Molchanov P, Hilliges O and Kautz J. 2020. Weakly supervised 3D hand pose estimation via biomechanical constraints//Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK: Springer: 211-228 [DOI: 10.1007/978-3-

030-58520-4_13］

Wen Y，Pan H，Yang L，Pan J，Komura T and Wang W. 2023. Hierarchical temporal transformer for 3D hand pose estimation and action recognition from egocentric RGB videos//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver，Canada：IEEE：21243-21253

Yang L X，Li K L，Zhan X Y，Lyu J，Xu W Q，Li J F，et al. 2022. ArtiBoost：boosting articulated 3D hand-object pose estimation via online exploration and synthesis//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. New Orleans，USA：IEEE：2740-2750［DOI：10.1109/CVPR52688.2022.00277］

Zheng C，Zhu S，Mendieta M，Yang T，Chen C and Ding Z. 2021. 3D human pose estimation with spatial and temporal transformers//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision（ICCV）. Montreal，Canada：IEEE Computer Society：11656-11665

Zhou Z，Zhou S，Lv Z，Zou M，Tang Y and Liang J. 2024. A simple baseline for efficient hand mesh reconstruction//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. Seattle，USA：IEEE Compter Society：1367-1376

作者简介

贾迪，男，教授，主要研究方向为摄影测量、三维视觉与空间感知、多传感器融合和机器人控制。
E-mail：lntu_jiadi@163.com
王建淳，通信作者，男，硕士研究生，主要研究方向为手势姿态估计和三维重建。E-mail：lntu_wangjianchun@163.com
韩雪峰，男，研究员，主要研究方向为人工智能和智慧矿山。
E-mail：173657676@qq.com
张藩，男，硕士研究生，主要研究方向为视觉基础模型和立体匹配。E-mail：lntu_zhangfan@163.com
王骁，男，硕士研究生，主要研究方向为语义分割。
E-mail：lntu_wangxiao@163.com