文章编号　1004-924X(2026)09-1411-12

复杂关节物体六自由度位姿估计与实时系统

欧林林[1*]，陈　婷[1]，禹鑫燚[1]，Umarov Tilek Mutalibovich[2]，姜浩男[1]
(1. 浙江工业大学 信息工程学院，浙江 杭州 310023；
2. 奥什国立大学，奥什 723500)

摘要：针对复杂关节物体六自由度位姿估计中存在的运动学约束建模不足等问题，提出了一种关节感知解耦的位姿估计方法。采用关节感知的解耦建模策略，利用独立分支分别回归物体旋转与平移，从而缓解二者之间的误差耦合。在此基础上，设计注意力引导的多模态特征融合机制，动态建模图像语义与点云几何间的跨模态相关性，实现遮挡稳定的特征提取。最后，设计自适应加权方案自动平衡多项损失函数，并结合轻量化预测模块实现部件铰接参数的回归。在 ArtImage 基准数据集上的实验结果表明，该方法在无需额外优化步骤的情况下，单张图像的推理速度最快可达 20 ms；对于抽屉等复杂多部件物体，旋转误差稳定在 1.6°；眼镜类别的 3D IoU 提升约 28%，在绝大多数类别和指标上均优于对比基准。该框架实现了从基座视角到各部件运动学的统一估计，在保证物理一致性的同时，显著提升了位姿预测的实时性与精确度，具备嵌入机器人实时视觉系统的应用潜力。

关　键　词：复杂关节物体；六自由度位姿估计；关节感知解耦；跨模态融合；自适应加权
中图分类号：TP391.41　　**文献标识码**：A
doi：10.37188/OPE.20263409.1411　　**CSTR**：32169.14.OPE.20263409.1411

6-DoF pose estimation and real-time system for complex articulated objects

OU Linlin[1*], CHEN Ting[1], YU Xinyi[1], Umarov Tilek MUTALIBOVICH[2], JIANG Haonan[1]

(1. *College of Information Engineering, Zhejiang University of Technology, Hangzhou* 310023, *China*;
2. *Osh State University, Osh* 723500, *Kyrgyzstan*)
* *Corresponding author, E-mail: linlinou@zjut.edu.cn*

Abstract: To address limitations in kinematic modeling for 6-DoF pose estimation of complex articulated objects, a joint-aware decoupled pose estimation framework is proposed. A decoupled modeling strategy is adopted, in which rotation and translation are regressed through independent branches, thereby reducing error coupling. An attention-guided multimodal feature fusion mechanism is developed to capture the relationships between image semantics and point cloud geometry, enhancing robustness under occlusion. An adaptive weighting scheme is further introduced to balance multiple loss terms during training. In addition, a lightweight module is designed to predict part-level articulation parameters. Experimental results on the ArtImage dataset demonstrate that the proposed method achieves an inference speed of up to 20 ms per

收稿日期：2026-02-13；**修订日期**：2026-03-23.
基金项目：国家自然科学基金资助项目(No. 62373329)；浙江省自然科学基金资助项目(No. LZ25F030003，No. LBMHD24F030002)

frame without requiring external optimization. For complex objects such as drawers, the rotation error remains stable at 1.6°, while the 3D IoU for the glasses category improves by approximately 28%. The method consistently outperforms baseline approaches across most categories and evaluation metrics. The proposed framework enables unified estimation from the base pose to articulated parts, improving both pose accuracy and inference efficiency while preserving physical consistency.

Key words: complex articulated objects; 6-DoF pose estimation; joint-aware decoupling; cross-modal fusion; adaptive weighting

1 引　言

部署于家庭和仓库环境的机器人需要具备操作抽屉、门、盖子及各类带关节结构物体的能力。此类物体的部件通常由平移关节或旋转关节连接，其运动轨迹受限于特定的运动学约束，而非单一固定的位姿。若缺乏各部件的六自由度(Six-Degree-Of-Freedom，6-DoF)位姿、关节轴线及运动范围限制信息，规划器将无法确定抓取位置、拉动方向或停止时机，从而导致操作失败甚至硬件损坏。在机器人操作任务中，铰接结构物体广泛存在，既包括家庭场景中的抽屉、门、盖子等，也包括仓储与工业环境中的柜体、机械连杆及装配结构等。近年来，关节物体位姿估计问题逐渐受到关注，相关研究主要围绕部件级规范表示、运动学约束建模以及遮挡与歧义问题的稳定处理展开。

与可视为三维空间单一实体的刚性物体不同，关节物体由多个通过关节连接的刚性部件组成，并受特定运动学的约束。Rehg等[1]的研究表明，关节物体运动学固有的自遮挡特性，因此需要基于运动学模型预测遮挡顺序。Li等[2]通过引入A-NCSH规范化表示和运动学约束，提升了部件位姿估计的稳定性，使模型在部分遮挡的观测条件下仍能保持较高的估计精度。为缓解对称性和自遮挡导致的歧义问题，Manhardt等[3]提出了多假设预测方法，通过分析假设分布检测歧义并提升位姿估计的稳定性。估计关节物体各刚性部件的6-DoF位姿(旋转和平移)已成为计算机视觉领域的重要课题，广泛应用于机器人控制、增强现实[4]及三维场景重建[5]等领域。

通过引入部件级规范表示，Li等[2]建立了可泛化至未知实例的关节物体位姿估计范式。国内研究人员(如Chen等[6])在刚性物体领域进一步探索了规范形状空间以提升泛化能力；同时，在六自由度位姿估计及三维重构方面也取得了一定进展[7-8]。然而，现有方法多数[9]以部件为中心建模，将关节物体视为多个刚性部件的简单组合，未充分考虑运动学结构的约束，导致部件间运动学差异被忽略；部分方法依赖预定义的关节标注或计算机辅助设计(Computer-Aided Design，CAD)模型，难以泛化至未知类别及复杂场景[10]。无CAD的类别级位姿估计通过特征匹配与结构重建[11]得以实现。此外，在剧烈铰接运动或严重遮挡下，当前方法存在预测不稳定、稳定性不足的问题。除利用形变来建模局部运动外[12-13]，多数工作侧重于全局位姿估计，缺乏对局部部件位姿的精细化建模。

关节物体本质上可分为运动不受约束的自由部件和运动被约束在固定关节轴上的受限部件。Li等基于ANCSH规范表示建立了自由-受限部件范式。Xue等[14]通过OMAD表示将这一思路扩展至形状几何变形与姿态变形的联合建模。Liu等[15]进一步提出ArtPERL，将复杂关节物体位姿估计转化为强化学习问题，通过以关节为中心的策略优化关节状态估计。Yu等[16]提出一种端到端框架，实现了自由部件特殊欧几里得群(Special Euclidean Group，SE(3))估计与受限部件状态回归的统一。上述部件规范的类别级表示方法在国内也得到深入研究。为进一步提升精度并缓解单阶段流水线中持续存在的旋转平移误差耦合，本文提出一种关节感知的解耦框架，可在无需额外优化步骤的条件下实现类别级6-DoF位姿预测。

该框架的核心在于以运动学为中心的设计，将基座位姿估计SE(3)目标解耦为两个协同的

子任务：旋转分量在特殊正交群（Special Orthogonal Group，SO(3)）上通过测地线监督回归，平移分量则在三维欧氏空间$\mathbf{R}^3$中利用部件感知的倒角（Chamfer）距离进行优化。两条分支共享深层编码器，在促进特征协同学习的同时防止跨域误差放大。针对自遮挡问题，设计注意力引导融合模块以动态建模图像语义与点云几何间的跨模态关联。此外，设计自适应不确定性重加权方案，无需人工调参即可自动平衡几何流形一致性、离群值稳定性与局部形状保真度。最后，通过轻量化关节状态预测模块回归所有受限连杆的轴角参数（Axis-Angle），并结合显式运动学链组合计算各连杆位姿，实现统一的关节物体位姿估计。

2　问题描述

复杂关节物体位姿估计的任务目标是：在已知物体类别先验的条件下，从单帧红绿蓝-深度（Red-Green-Blue-Depth，RGB-D）图像中恢复该物体在相机坐标系下的全局位姿、内部各部件的相对运动状态以及几何参数。

2.1　位姿表示与坐标系定义

设相机坐标系为$\{C\}$。对于某一类关节物体，定义其规范空间下的模型为$\{M\}$。该物体由一个基座部件和$K-1$个铰接部件组成。基座部件在相机坐标系下的位姿由变换矩阵$T_{C\leftarrow B}\in \mathrm{SE}(3)$表示：

$$T_{C\leftarrow B}=\begin{bmatrix} R_{C\leftarrow B} & t_{C\leftarrow B} \\ 0^{\mathrm{T}} & 1 \end{bmatrix}, \tag{1}$$

其中：旋转矩阵$R_{C\leftarrow B}\in \mathrm{SO}(3)$描述空间朝向，平移向量$t_{C\leftarrow B}\in \mathbf{R}^3$描述空间位置。

2.2　铰接运动学建模

第k个铰接部件相对于基座的运动受限于特定的物理关节（如转动副或移动副）。本文将部件k的位姿分解为相对于关节中心$\{J_k\}$的局部变换。部件k在相机坐标系下的最终位姿$T_{C\leftarrow k}$可表示为以下运动学复合形式：

$$T_{C\leftarrow k}=T_{C\leftarrow B}\cdot T_{B\leftarrow J_k}\cdot T_{J_k}(\theta_k), \tag{2}$$

其中：$T_{B\leftarrow J_k}$为关节中心在基座坐标系下的静态几何属性（包括关节轴线方向和位置），θ_k为标量关节状态（转动角度或移动距离），$T_{J_k}(\theta_k)$为由该状态产生的动态变换矩阵。

2.3　旋转与平移的耦合问题

在传统的位姿回归任务中，旋转误差往往会通过力臂效应传播至平移分量，导致铰接部件的对齐失效。本文的目标是构建一个映射函数Φ，通过从输入模态中提取解耦特征，实现对旋转矩阵R与平移残差Δt的独立推理，从而缓解误差耦合：

$$\{R_{C\leftarrow B},t_{C\leftarrow B},\theta_{1:K-1}\}=\Phi(I_{\mathrm{RGB}},D_{\mathrm{depth}},P_{\mathrm{prior}}), \tag{3}$$

其中：$\theta_{1:K-1}$表示物体内部$K-1$个关节的状态集合（如转动角度或移动位移），I_{RGB}为输入的红绿蓝（Red-Green-Blue，RGB）彩色图像，D_{depth}为对应的深度图，P_{prior}为从模型库中提取的类别级形状先验信息，Φ需在保证SO(3)流形一致性的前提下，实现端到端的实时推理。

3　关节感知的解耦位姿估计框架

本文提出了一种端到端的复杂关节物体6D位姿估计框架。该框架的核心设计包括：(1)关节感知的解耦建模策略，有效缓解旋转与平移间的误差耦合；(2)注意力引导的跨模态融合机制，在遮挡环境下校准跨模态特征；(3)组合几何对齐损失函数，通过测地线距离、部件感知Chamfer距离及Huber项约束流形一致性；(4)基于残差细化与关节状态建模的跨关节策略，通过残差细化提升旋转预测精度，并结合显式运动学链组合计算各部件位姿。

该方法的具体流程如图1所示。首先，将输入的RGB图像、由深度图反投影得到的点云及类别形状先验分别通过残差网络ResNet-50[17]、分层点云处理网络PointNet++[18]和先验编码器提取特征，并通过挤压-激励（Squeeze-and-Excitation，SE）[19]模块对各模态特征进行初步增强。随后，将各模态特征映射到统一维度后拼接形成768维融合输入，并送入多模态特征融合模块进行自适应特征整合，从而生成融合特征表示。该特征随后被分别送入解耦的六维旋转分支与三维平移分支，以预测基座的6D位姿。同时利用HS-Encoder[20]模块对规范化部件点云及关节参数进行编码嵌入。最后，差分状态网络（State Network）[16]输出关节状态，并通过显式运动学链组合将基座位姿映射为各连杆的6D位姿。整个框架通过Chamfer损失[21]、角度损失及运动学损

失进行端到端训练，推理过程无需额外的非可微后处理优化。

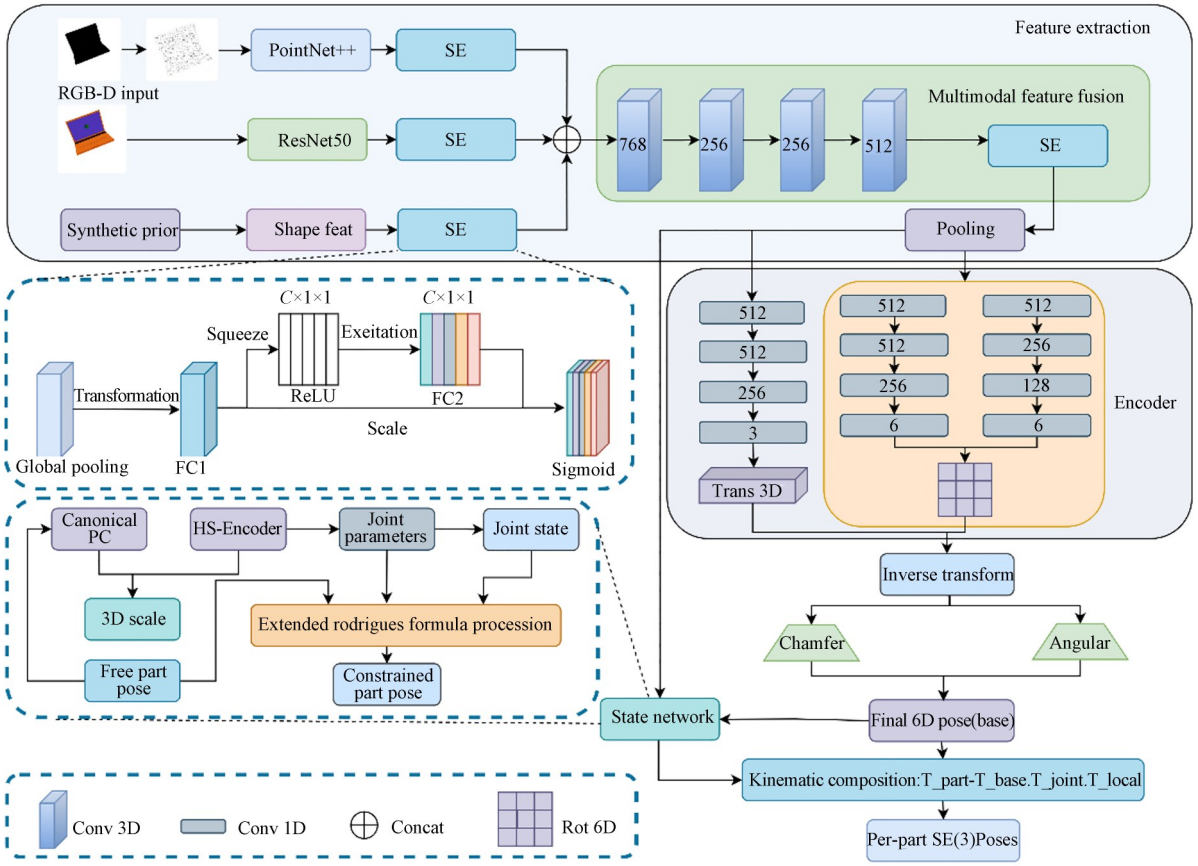

图 1　关节感知解耦网络

Fig. 1　Joint-aware decoupled network

3.1　旋转与平移的关节感知解耦建模

为缓解铰接部件运动时旋转和平移之间的相互漂移，设计了一种关节感知的解耦建模策略，通过两个独立且协同演化的分支分别回归旋转和平移，缓解误差耦合。该架构摒弃了单一的 SE(3) 预测头，转而配置双分支结构：(1) 由 SO(3) 上的测地线损失监督的旋转分支；(2) 预测相对于关节中心先验的残差偏移的平移分支。旋转与平移分支共享编码器提取的特征，但保持独立的批归一化(Batch Normalization，BN)与 Dropout 统计信息。该设计实现了特定任务的特征对齐，确保旋转分支的角度噪声不会传播至平移预测任务中。此外，该架构允许网络直接学习连续的几何流形表示，无需通过奇异值分解(Singular Value Decomposition，SVD)进行后期正交性修正，从而使网络能够捕捉在大幅度关节运动下依然稳健的解耦特征。

在建模方式上，各活动部件的位姿并非直接在相机坐标系下预测，而是定义在相对于铰接关节的局部坐标系中。从关节坐标系到相机坐标系的变换矩阵可表示为：

$$T_{C\leftarrow J}=\begin{bmatrix}\boldsymbol{R}_{\text{final}} & \boldsymbol{t}_{\text{final}}\\ 0^{\mathrm{T}} & 1\end{bmatrix}, \tag{4}$$

其中：旋转矩阵 $\boldsymbol{R}_{\text{final}}$ 与平移向量 $\boldsymbol{t}_{\text{final}}$ 分别定义为：

$$\begin{cases}\boldsymbol{R}_{\text{final}}=\boldsymbol{R}_{C\leftarrow B}(\boldsymbol{R}_{B\leftarrow J})\\ \boldsymbol{t}_{\text{final}}=\boldsymbol{R}_{C\leftarrow B}\boldsymbol{t}_{B\leftarrow J}+\boldsymbol{t}_{C\leftarrow B}\end{cases}, \tag{5}$$

其中：$\boldsymbol{R}_{C\leftarrow B}$ 与 $\boldsymbol{t}_{C\leftarrow B}$ 表示基座位姿，$\boldsymbol{R}_{B\leftarrow J}$ 与 $\boldsymbol{t}_{B\leftarrow J}$ 编码与关节相关的运动。这种层次化分解确保了局部部件运动与铰接结构的一致性，同时允许网络回归具有物理可解释性的参数。

3.1.1　旋转分支

采用连续的 6D 表示法，以避免欧拉角或四元数固有的不连续性和奇异性。网络预测两个三维向量 $u,v\in\mathbf{R}^3$，并通过施密特(Gram-

Schmidt)正交化进行处理：

$$\begin{cases} \boldsymbol{r}_1 = \dfrac{\boldsymbol{u}}{\|\boldsymbol{u}\|} \\ \boldsymbol{r}_2 = \dfrac{\boldsymbol{v} - (\boldsymbol{r}_1^{\mathrm{T}}\boldsymbol{v})\boldsymbol{r}_1}{\|\boldsymbol{v} - (\boldsymbol{r}_1^{\mathrm{T}}\boldsymbol{v})\boldsymbol{r}_1\|} \\ \boldsymbol{r}_3 = \boldsymbol{r}_1 \times \boldsymbol{r}_2 \end{cases}. \tag{6}$$

这些基向量构成了一个有效的旋转矩阵 $\boldsymbol{R} = [\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3] \in \mathrm{SO}(3)$。关键之处在于微分图中保留了正交化过程，使梯度能够直接作用于原始6D表示，无需依赖SVD进行后处理投影。该分支由SO(3)上的测地线损失监督，确保部件旋转接近180°时仍能保持稳定的梯度更新。

3.1.2　平移分支

平移分支独立预测一个三维偏移向量 $\boldsymbol{t} \in \mathbf{R}^3$。最终的变换矩阵表示为：

$$T = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{t} \\ 0^{\mathrm{T}} & 1 \end{bmatrix}. \tag{7}$$

与传统的联合回归不同，此处的平移预测与旋转解耦确保角度噪声不会导致位置漂移。这种独立性对于铰接运动至关重要，因为局部坐标系原点(即关节中心)会随关节角的变化而移动。此外，最终偏移向量由HS-Encoder提供的关节中心先验与经学习参数缩放的残差位移相加得到，使网络在共享旋转分支特征的同时能够自适应调节步长。

3.1.3　几何与物理意义

该关节感知架构具备两个核心优势。首先，体现为误差隔离：旋转和平移分别由独立的分支处理，梯度回传时各参数更新互不干扰，因此训练过程更为稳定。相较而言，联合估计易使朝向偏差与位置偏移相互耦合，造成双向精度损失。其次，表现为运动学一致性：以关节局部坐标系为基准构建位姿表征，并结合显式运动学链组合将基座位姿与关节状态映射为各部件位姿，使所得变换与转动或平移约束保持一致，从而避免额外的非可微运动学校正步骤。定量评估显示，上述解耦策略改善了大角度旋转场景的稳定性，同时对局部遮挡具有更好的稳定性。模型倾向于以关节中心进行几何推理，而非依赖全局坐标偏移，此种行为模式与关节物体的实际物理特性更为吻合。

3.2　注意力引导的多模态特征融合机制

为充分挖掘RGB图像、点云以及类别先验的互补优势，本文设计了一种由多层感知机(MLP)与挤压-激励(SE)通道注意力机制组成的轻量级多模态融合模块。该模块先对拼接特征进行非线性映射，再进行通道重标定，实现跨模态信息的自适应融合，并为下游位姿预测提供具有遮挡稳定性的特征表示。在具体实现中，拼接后的768维特征首先通过多层感知机进行维度映射与非线性融合(768→256→256→512)，随后输入SE模块进行通道注意力加权，使网络能够根据不同模态的信息自适应调整通道响应。当深度几何信息较为准确时，网络倾向于抑制冗余纹理通道；当纹理信息较弱时，则增强几何相关通道的响应。

具体而言，视觉流通过卷积编码器(如残差网络ResNet-50)提取语义特征表示为：

$$A \in \mathbf{R}^{C_a}. \tag{8}$$

几何流利用分层点云处理网络(hierarchical PointNet++)对采样点云进行编码，得到几何特征表示为：

$$G \in \mathbf{R}^{C_g}. \tag{9}$$

先验流则对类别级的形状信息进行编码，得到先验特征表示：

$$P \in \mathbf{R}^{C_p}. \tag{10}$$

上述各模态特征表示通过可学习的线性变换投影至统一的隐空间维度，随后进行拼接以构建统一表示：

$$\boldsymbol{Z} = [\boldsymbol{W}_a \boldsymbol{A} \,\|\, \boldsymbol{W}_g \boldsymbol{G} \,\|\, \boldsymbol{W}_p \boldsymbol{P}] \in \mathbf{R}^{C}, \tag{11}$$

其中：$\boldsymbol{W}_a$，$\boldsymbol{W}_g$和$\boldsymbol{W}_p$分别为对应模态的投影矩阵。尽管直接拼接融合了各模态信息，但其对所有通道一视同仁，无法根据不同样本动态调整模态重要性。为克服这一限制，本文引入SE通道注意力机制，根据全局相关性动态重构特征通道权重。给定融合特征张量$\boldsymbol{Z}$，其注意力权重α计算如下：

$$\alpha = \sigma(\boldsymbol{W}_2 \delta(\boldsymbol{W}_1 \mathrm{GAP}(\boldsymbol{Z}))), \tag{12}$$

其中：GAP(·)表示全局平均池化，$\delta(\cdot)$为ReLU激活函数，$\sigma(\cdot)$为Sigmoid函数，矩阵$\boldsymbol{W}_1$和$\boldsymbol{W}_2$分别执行通道压缩与扩张，以极低的计算开销建模通道间的依赖关系。最终的精炼特征通过元素级乘法获得：

$$\tilde{\boldsymbol{Z}} = \alpha \odot \boldsymbol{Z}. \tag{13}$$

这种注意力引导的融合机制可选择性强化与铰接部件及判别性几何线索相关的通道响应，同时抑制背景噪声与冗余特征。因此，网络无需

额外监督即可自适应地平衡视觉、几何与先验信息，为下游位姿估计构建稳健且具强表达能力的特征空间。

该SE门控融合机制能够有效提升旋转一致性与几何对齐精度，尤其在遮挡或纹理缺失场景下表现出更好的稳定性。相较于多头注意力机制，本设计专注于通道间的重要性权重而无需显式的空间交互，从而形成了一个高度轻量化的模块。该模块可无缝集成至现有的RGB-D流线中，且无需更改架构即可扩展至更多模态。通过在统一框架内聚合跨模态线索，本文方法实现了关节物体可靠且物理一致的位姿预测。

3.3 耦合流形约束与几何对齐的复合损失函数

损失函数的设计直接影响模型的训练稳定性与最终精度。传统的基于L_2范数的损失函数对离群点过度敏感，且未能遵循旋转矩阵的流形结构。为解决上述问题，本文设计并实现了一种三项式复合损失函数，该函数在流形上完全可微，且减少人工权重调参需求。

首先，针对旋转分支，本文采用定义在特殊正交群上的测地距离进行监督。给定真实旋转矩阵$\boldsymbol{R}_{\mathrm{gt}}$与预测旋转矩阵$\boldsymbol{R}_{\mathrm{pred}}$，旋转损失函数定义为：

$$\mathcal{L}_{\mathrm{rot}}=\arccos\left(\frac{\mathrm{Tr}(\boldsymbol{R}_{\mathrm{gt}}^{\mathrm{T}}\boldsymbol{R}_{\mathrm{pred}})-1}{2}\right),\tag{14}$$

其中：$\mathrm{Tr}(\cdot)$表示矩阵的迹。该公式直接度量了流形上两个旋转之间的最小角度距离，在保持平滑性的同时，避免了欧拉角或四元数表示中常见的间断性问题。通过在SO(3)流形上直接进行优化，网络能够学习并生成几何有效的旋转表征，即使在关节剧烈运动时仍能保持预测一致性。与现有方法在优化后才投影回SO(3)的框架不同，本文直接对原始矩阵输出进行监督，确保梯度在整个训练过程中始终约束于流形之上，从而无需奇异值分解(Singular Value Decomposition，SVD)修正。

对于平移预测，本文采用Smooth-L_1损失(即Huber损失)。该损失函数兼L_1的稳定性和L_2的可微性，定义如下：

$$\mathcal{L}_{\mathrm{trans}}=\begin{cases}\dfrac{1}{2\beta}\|\Delta t\|_2^2, \text{if } \|\Delta \mathrm{t}\|_2<\beta,\\ \|\Delta \mathrm{t}\|_2-\dfrac{1}{2}\beta, \text{otherwise.}\end{cases},\tag{15}$$

其中：β为控制过渡区域的阈值参数。本文将该损失封装为单一的自动微分函数，使离群点的权重在较大偏差范围内自动衰减，无需人工截断处理。为进一步确保精细的几何一致性，本文引入部件感知倒角距离，约束各铰接部件的预测点集与真实点集精确对齐。设$\mathcal{P}$和$\mathcal{Q}$分别代表预测与目标的部件表面点云，部件级倒角损失为：

$$\mathcal{L}_{\mathrm{chamfer}}=\frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}}\min_{q\in\mathcal{Q}}\|p-q\|_2^2+\frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}}\min_{p\in\mathcal{P}}\|p-q\|_2^2.\tag{16}$$

与全局倒角损失不同，本文逐连杆计算该损失项，确保薄壁结构(如铰链、把手)能被精确对齐，而不会被大体积主体表面所稀释。该损失作用于局部部件的对应关系，在保持表面连续性的同时，引导模型准确重建相对关节配置。实验表明，该设计显著提升了精细结构的对齐精度，且不会对全局形状先验产生过拟合。最终的总目标函数表示为：

$$\mathcal{L}_{\mathrm{total}}=\lambda_r\mathcal{L}_{\mathrm{rot}}+\lambda_t\mathcal{L}_{\mathrm{trans}}+\lambda_c\mathcal{L}_{\mathrm{chamfer}},\tag{17}$$

其中：$\lambda_r,\lambda_t,\lambda_c$为可学习的标量参数(参数初始化为1，并在训练过程中自适应更新)。在训练过程中与网络参数一同通过反向传播进行优化，用于自适应平衡旋转、平移及局部几何对齐三项损失的贡献。该设计避免了人工设定权重带来的调参困难，并能够根据不同训练阶段的优化需求动态调整各损失项的重要性，从而有助于提升整体训练的稳定性与收敛性能。在实际训练中，该参数在标准优化策略下稳定收敛，未引入额外的训练不稳定性。这种复合损失设计同时约束流形一致的旋转估计、稳健的平移回归以及精确的局部几何对齐。因此，模型在未见过的关节配置下表现出较好的泛化能力，训练数值稳定性得以增强，且在真实场景中的关节物体位姿估计精度有效提升。

3.4 基于残差细化与潜状态估计的跨关节策略

在关节物体位姿估计中，单阶段旋转回归往往难以实现精确对齐，尤其是在物体发生大幅度旋转或涉及多关节耦合时。从高维几何与语义特征直接回归位姿往往只能得到粗略朝向估计，无法满足细粒度角度精度的需求。为克服此局限，本文提出分阶段旋转细化框架，通过渐进式残差误差修正，辅以转角增强策略来提升模型的

训练泛化能力。

3.4.1　分阶段旋转细化

本文在旋转分支中引入分阶段细化策略，即通过粗略预测与残差修正的两步方式逐步优化旋转估计。需要说明的是，该细化过程完全在网络前向传播中实现，属于模型内部的结构设计，不改变整体框架的端到端特性。

为了修正残差误差，细化头部预测一个三维残差向量 $\Delta \boldsymbol{r} \in \mathbf{R}^3$，该向量表示定义在 so(3) 切空间中的轴角扰动。利用向量空间 $\mathbf{R}^3$ 与李代数 so(3) 的同构关系，$\Delta \boldsymbol{r}$ 可映射为反对称矩阵 $[\Delta \boldsymbol{r}]_\times \in \mathrm{so}(3)$。在李群框架下，旋转更新通过指数映射 $\exp: \mathrm{so}(3) \to \mathrm{SO}(3)$ 实现，从而得到细化后的旋转矩阵：

$$\boldsymbol{R}_{\mathrm{ref}} = \exp\left([\Delta \boldsymbol{r}]_\times\right) \boldsymbol{R}_0, \tag{18}$$

其中：$[\cdot]_\times$ 表示三维向量对应的反对称矩阵。该更新形式对应于在旋转流形上的左乘扰动，使残差修正始终约束在 SO(3) 上，从而保持旋转矩阵的正交性与连续性。在反向传播过程中，梯度通过指数映射在李群上进行传递，实现粗略阶段与细化阶段的端到端优化。

当残差扰动较小时，指数映射可进行一阶近似：

$$\exp([\Delta \boldsymbol{r}]_\times) \approx \boldsymbol{I} + [\Delta \boldsymbol{r}]_\times. \tag{19}$$

因此，该更新在局部可视为线性化的李代数扰动，具有良好的数值稳定性。为进一步控制高阶非线性误差，本文约束 $\| \Delta \boldsymbol{r} \| \leqslant \pi/6$，从而避免大角度更新带来的误差积累。实验表明，这种层次化方法有效减少了取向误差的累积，尤其适用于剪刀、眼镜等具有多个运动部件且存在旋转耦合的复杂物体。

3.4.2　面向关节多样性的转角增强

为了提升模型在不同关节状态下的泛化性能，本文在训练阶段引入了转角增强策略，如图 2 所示。对于每个关节参数真实值 θ_{gt}，本文在限定范围内注入随机高斯扰动：

$$\tilde{\theta} = \theta_{\mathrm{gt}} + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2), \tag{20}$$

其中：σ 控制扰动强度，$\tilde{\theta}$ 表示预测角度，θ_{gt} 为真实角度，ϵ 为服从高斯分布的噪声项。随后，重新计算对应的真实旋转矩阵：

$$\boldsymbol{R}_{\mathrm{aug}} = \boldsymbol{R}_{\mathrm{axis}}(\tilde{\theta}), \tag{21}$$

其中：$\boldsymbol{R}_{\mathrm{axis}}(\cdot)$ 表示绕预定义铰接轴生成的旋转矩阵。这种增强手段有效拓宽了观测位姿的关节分布，使模型能够接触到未见过的角度配置，从而显著增强了模型对关节极限变化及遮挡场景的稳定性。

3.4.3　潜状态网络

本文集成并改进了潜状态网络，该架构参考 EfficientCAPER 的设计，采用 HS-Encoder 分支。该分支提取规范化部件点云及当前三维尺度估计特征，将其编码为部件感知潜向量，并与预测的关节参数（包括关节轴向与相对原点位置）拼接。随后，通过扩展罗德里格斯公式（Extended rodrigues-formula）处理器将关节参数向

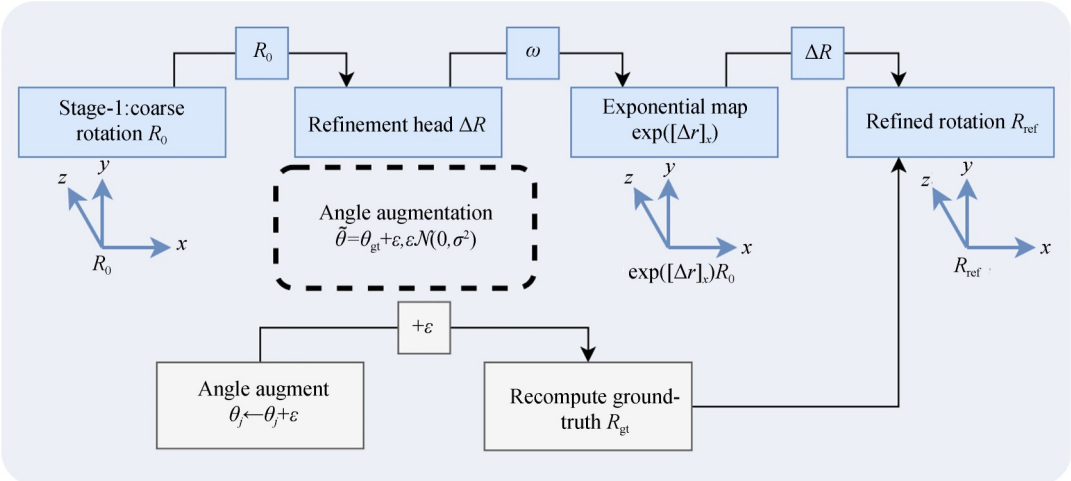

图 2　旋转细化与角度增强

Fig. 2　Rotation refinement and angle augmentation

量转换为描述受限部件与其父节点之间相对关系的瞬时旋转与平移。基于上述关节状态表示，结合基座预测头输出的自由部件位姿，通过显式运动学链组合计算各受限部件的最终位姿，从而保证结果满足关节运动约束。该过程可在单次前向传播中完成，无需额外的非可微优化步骤。

尽管精准的基础位姿估计提供了稳定的全局参考，但若不模拟个体部件的运动，则无法完整描述关节物体。为此，本文整合该状态网络来预测关节状态参数，并结合全局基础位姿通过显式运动学链组合计算部件级位姿。给定输入特征F，状态网络回归关节角度为：

$$\theta = f_{\text{state}}(F), \tag{22}$$

结合全局基础位姿T_{base}，最终的部件位姿计算为：

$$T_{\text{part}} = T_{\text{base}} \cdot T(\theta), \tag{23}$$

其中：$T(\theta)$是由关节角参数化的变换矩阵。这种分层预测策略在确保基础稳定性的同时，实现了对部件级运动的灵活建模，有效减少了关节耦合导致的误差，使得框架在处理复杂铰接结构时更具可扩展性与推理效率。

4 实验结果及分析

4.1 实验设置

本研究在ArtImage数据集上进行算法性能评估。该数据集是一个专门针对关节物体位姿估计设计的合成基准，其三维模型源自PartNet-Mobility库[22]。ArtImage数据集涵盖了笔记本电脑、眼镜、洗碗机、剪刀和抽屉五类典型关节物体。每类物体均包含多种关节运动状态，并提供基座位姿及部件级变换的真值标注。由于其复杂的关节配置，该数据集是验证关节感知建模有效性的理想选择。

为全面评价算法性能，本文采用以下四项指标：旋转误差：预测旋转矩阵与真实值之间的测地线距离，用于衡量角度精度；平移误差：预测平移向量与真实值之间的欧氏距离，反映空间位移精度；3D IoU(%)：预测三维边界框与真实框的交并比，评估尺度估计及整体空间对齐效果；推理时间：测试阶段单张图像的平均运行时间，衡量算法的计算效率。

在数据预处理阶段，原始深度图被反投影至三维空间并转换为点云。为平衡计算效率与几何保真度，每个点云均被统一降采样至1 024个点。本文方法基于PyTorch框架实现，并在NVIDIA RTX 3090 GPU上进行训练。网络采用AdamW优化器进行优化，初始学习率设为8×10^{-4}，权重衰减为1×10^{-4}，Batch Size设为8。学习率采用StepLR策略(Step size=5，Gamma=0.5)。模型共训练250个Epoch，并根据验证集准确率选取最优模型参数。

4.2 定量评价与对比

在ArtImage数据集上进行了定量评估，并与现有主流方法进行了对比分析。表1汇总了五类物体的旋转误差、平移误差、3D IoU及推理速度表现。旋转/平移误差及推理时间数值越低越好，3D IoU数值越高越好。

如表1所示，本文方法在绝大多数类别和指标上均优于对比基准。关节感知的解耦建模显著降低了旋转和平移误差；而引入注意力机制的多尺度特征融合则有效捕获了局部几何信息，提升了3D IoU表现。此外，分阶段细化策略确保了在大角度铰接情况下的预测稳定性。值得注意的是，本模型推理速度极快，单张图像处理最快仅需约20 ms，展现出较好的实时性能。为进一步印证定量分析结果，图3展示了定性视觉效果。这些可视化算例直观地揭示了不同设计决策对基座定位、旋转稳定性以及铰接部件运动一致性的影响。视觉证据与数值结果高度契合，有力证明了本文框架在处理挑战性案例时的优越性。

4.3 消融实验

为分析各核心组件对系统性能的贡献，本文在洗碗机上进行了消融实验。该类别包含典型的铰接结构，几何部件较为丰富，能够在一定程度上反映模型在常见关节物体上的行为特征。本文在主实验中已对多个类别(如抽屉、剪刀、眼镜等)进行了全面评估，从整体结果上验证了所提方法在不同运动学结构下的有效性。所有变体模型均在相同的数据划分、优化器设置及增强策略下训练30个Epoch，并统一采用1 024个采样点以确保实验公平性。

表 1　估计精度与实时性对比

Tab. 1　Comparison of pose accuracy and runtime performance

Category	Method	Per-part Pose			Inference time per image/s
		Rotation error/(°)	Translation error/m	3D IOU/%	
Laptop	A-NCSH[2]	5.3,5.4	0.054,0.043	56.7,40.2	9.0
	OMAD[12]	5.4,**4.3**	0.062,0.061	43.5,24.1	1.6
	ArtPERL[13]	4.9,4.7	0.053,0.066	64.6,50.4	0.9
	Our-method	**3.09**,5.82	**0.038,0.038**	**73.8,73.4**	**0.021**
Eyeglasses	A-NCSH[2]	**3.7**,22.3,23.2	0.049,0.313,0.324	52.5,40.2,39.6	11.9
	OMAD[12]	4.9,7.5,7.5	0.062,0.103,0.324	22.8,20.5,21.4	2.5
	ArtPERL[13]	4.1,**6.2,6.0**	**0.047**,0.095,0.091	58.6,46.5,51.7	1.0
	Our-method	7.7,8.9,8.6	0.063,**0.092,0.091**	**86.8,83.8,83.9**	**0.186**
Dishwasher	A-NCSH[2]	4.0,4.8	0.059,0.123	84.3,56.2	5.5
	OMAD[12]	6.0,6.2	0.104,0.142	66.5,38.9	1.6
	ArtPERL[13]	3.9,4.3	0.055,**0.079**	89.3,67.6	0.9
	Our-method	**2.47,3.01**	**0.053**,0.091	**90.67,83.01**	**0.11**
Scissors	A-NCSH[2]	**2.0**,2.9	0.035,0.025	46.5,44.8	6.5
	OMAD[12]	3.9,3.4	0.048,0.039	35.6,34.5	1.7
	ArtPERL[13]	2.2,**2.6**	**0.031**,0.042	40.9,46.3	0.8
	Our-method	5.0,5.2	0.048,**0.010**	**73.9,69.1**	**0.020**
Drawer	A-NCSH[2]	2.8,3.5,3.9,2.9	**0.045**,0.155,0.157,**0.075**	90.2,81.5,78.4,82.7	16.5
	OMAD[12]	4.4,4.4,4.4,4.4	0.111,0.143,0.144,0.115	75.8,73.4,70.2,71.3	1.9
	ArtPERL[13]	3.5,3.5,3.5,3.5	0.061,0.112,0.121,0.104	84.8,78.6,79.0,81.2	1.1
	Our-method	**1.6,1.6,1.6,1.6**	0.047,**0.103,0.101**,0.081	**92.1,85.2,85.3,88.1**	**0.124**

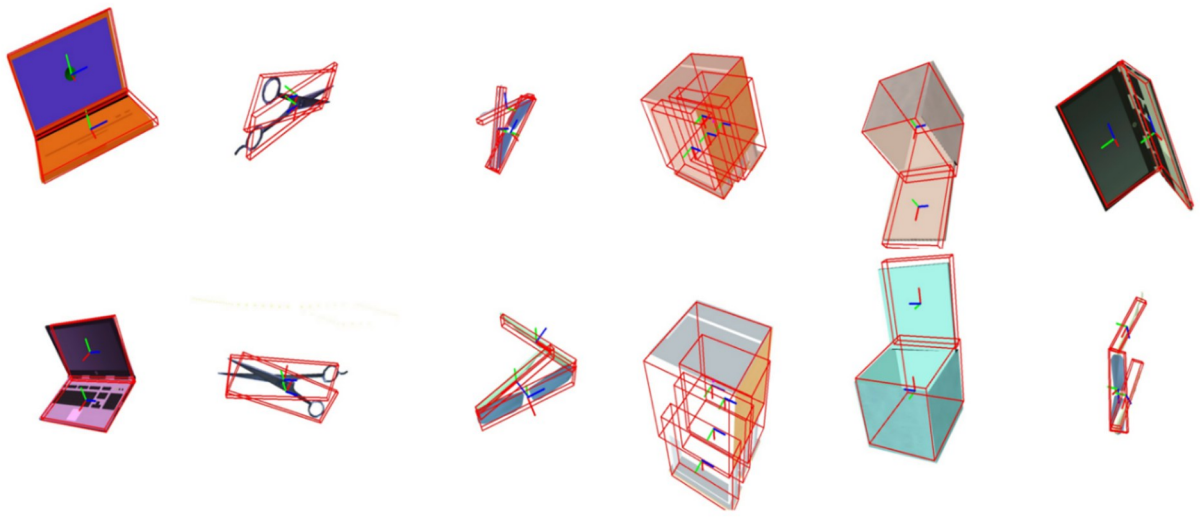

图 3　定性可视化结果

Fig. 3　Qualitative visualization results

对比变体包括：Full（本文完整模型）：包含所有设计组件；w/o decoupling：移除解耦头部，直接联合回归旋转 $\boldsymbol{R}$ 与平移 $\boldsymbol{t}$；w/o attention fusion：保留 RGB 与点云融合，但移除注意力机制（采用直接拼接）；w/o two-stage refinement：丢弃旋转细化头部；w/o part-aware chamfer：移除损

表 2　消融实验结果

Tab. 2　Results of ablation study

Variant	Rotation error/(°)	Translation error/m	3D IOU/%	Inference time per image/s
Full (ours)	2.5	0.038	77.8	0.023
w/o decoupling	2.7	0.061	62.3	0.019
w/o attention fusion	3.8	0.0594	55.1	0.021
w/o two-stage refinement	7.6	0.046	68.9	0.019
w/o part-aware chamfer	3.1	0.043	41.2	0.018

失函数中的部件感知几何项。

表 2 结果显示，移除解耦模块后，平移误差显著增大，这证实了联合回归会导致旋转误差向平移预测传播（即误差耦合）。禁用注意力机制会导致 3D IoU 下降，且在复杂背景下的旋转预测精度受损，凸显了多模态特征选择性对齐的关键作用。缺乏分阶段细化时，大角度旋转案例出现了严重的尾部误差分布。移除部件感知 Chamfer 项会削弱模型对门边缘、把手等细微平面的几何一致性约束，导致 3D IoU 下降。综上所述，完整模型实现了精度与效率间的最优平衡。解耦机制抑制了误差传播，注意力融合确保了精确对齐，分阶段细化提升了极值稳定性，而改进的 Chamfer 损失则保障了部件层级的细节精度。

5　结　论

本文提出了一种复杂关节物体位姿估计框架，可在单次前向传播中同步恢复物体基座位姿与各部件的铰接参数。提出了一种关节感知的位姿解耦建模方法，通过在网络结构层面对旋转与平移进行分支建模并分别回归，有效缓解了关节物体运动过程中的误差耦合问题。设计了基于注意力门控的跨模态特征融合机制。该机制能够动态识别并增强对位姿估计具有高贡献度的关键通道特征，有效抑制了背景噪声。实验表明，该方法在细粒度类别及局部遮挡等挑战性场景下，依然展现出优秀的几何感知能力与稳定性。构建了端到端的轻量化联合估计架构。通过整合复合流形损失函数与潜状态预测器，该框架避免了对传统非可微后处理对齐步骤的依赖。在绝大多数类别和指标上均优于对比基准。同时，推理速度达到 20～186 ms/单张图像。证明了显式运动学建模与深度学习结合在实时感知任务中的高效性。实验表明，本框架在 ArtImage[15] 数据集上展现出优异的物理一致性：笔记本电脑、剪刀等类别的推理速度达 21 ms/单张图像（无需额外优化步骤），实时性能提升显著；对于抽屉等复杂多部件物体，旋转误差稳定在 1.6°，眼镜类别的 3D IoU 提升约 28%，充分验证了本文所提框架的优越性。

尽管本文方法在合成数据集上取得了优异表现，但目前仍面临虚实间隙带来的挑战。后续研究将重点关注以下内容：首先，采集包含复杂光照与真实传感器噪声的大规模实拍数据集，并结合领域自适应技术提升模型的跨域泛化能力；其次，引入不确定性量化机制为预测结果提供置信度度量，并探索基座与部件位姿间的双向反馈机制以进一步消除累积误差；最后，学习统一的跨类别铰接嵌入空间，以实现对未知类别的零样本迁移，进而将该框架部署于真实的机器人闭环操控系统中。

作者贡献声明：

欧林林：实验指导及论文审核；

陈婷：方法提出，实验设计及数据分析，论文构思与撰写；

禹鑫燚：论文审核与方向把控；

Umarov Tilek Mutalibovich：实验指导；

姜浩男：方法提出和设计。

参考文献:

[1] REHG J M, KANADE T. Model-based tracking of self-occluding articulated objects[C]. *Proceedings of IEEE International Conference on Computer Vision. June* 20-23, 1995, *Cambridge*, *MA*, *USA*. IEEE, 2002: 612-617.

[2] LI X L, WANG H, YI L, *et al*. Category-level articulated object pose estimation[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 13-19, 2020, *Seattle*, *WA*, *USA*. IEEE, 2020: 3703-3712.

[3] MANHARDT F, ARROYO D M, RUPPRECHT C, *et al*. Explaining the ambiguity of object detection and 6D pose from visual data[C]. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 27- *November* 2, 2019, *Seoul*, *Korea*. IEEE, 2020: 6840-6849.

[4] WENG Y J, WANG H, ZHOU Q, *et al*. CAPTRA: CAtegory-level pose tracking for rigid and articulated objects from point clouds[C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 10-17, 2021, *Montreal*, *QC*, *Canada*. IEEE, 2022: 13189-13198.

[5] LIN S J, FANG J D, IRSHAD M Z, *et al*. SplArt: articulation estimation and part-level reconstruction with 3D Gaussian splatting [EB/OL]. 2025: *arXiv*: 2506.03594. https://arxiv.org/abs/2506.03594

[6] CHEN D S, LI J, WANG Z, *et al*. Learning canonical shape space for category-level 6D object pose and size estimation[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 13-19, 2020. *Seattle*, *WA*, *USA*. IEEE, 2020: 11970-11979.

[7] 陈敏佳, 盖绍彦, 达飞鹏, 等. 采用辅助学习的物体六自由度位姿估计[J]. 光学 精密工程, 2024, 32(6): 901-914.
CHEN M J, GAI SH Y, DA F P, *et al*. Object 6-DoF pose estimation using auxiliary learning[J]. *Optics and Precision Engineering*, 2024, 32(6): 901-914. (in Chinese)

[8] 程瑶, 吴哲滔, 石肖伊, 等. 基于双目结构光周视扫描的3D件三维重构与位姿估计研究[J]. 光学 精密工程, 2025, 33(2): 337-347.
CHENG Y, WU ZH T, SHI X Y, *et al*. Research of 3D reconstruction and position estimation of 3D pieces based on binocular structured light peripheral scanning[J]. *Optics and Precision Engineering*, 2025, 33(2): 337-347. (in Chinese)

[9] HUO Y K, MENG X H, ZHANG L, *et al*. DiffArt: category-level articulation pose estimation *via* conditional diffusion[C]. 2025 *IEEE International Conference on Multimedia and Expo* (*ICME*). *June* 30-*July* 4, 2025, *Nantes*, *France*. IEEE, 2025: 1-6.

[10] WANG H, SRIDHAR S, HUANG J W, *et al*. Normalized object coordinate space for category-level 6D object pose and size estimation[C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 15-20, 2019, *Long Beach*, *CA*, *USA*. IEEE, 2020: 2637-2646.

[11] SUN J M, WANG Z H, ZHANG S Y, *et al*. OnePose: one-shot object pose estimation without CAD models[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 6815-6824.

[12] YI L, HUANG H B, LIU D F, *et al*. Deep part induction from articulated object pairs[J]. *ACM Transactions on Graphics*, 2018, 37(6): 1-15.

[13] YAN Z H, HU R Z, YAN X G, *et al*. RPM-Net: recurrent prediction of motion and parts from point cloud[J]. *ACM Transactions on Graphics*, 2019, 38(6): 1-15.

[14] XUE H, LIU L, XU W Q, *et al*. OMAD: object model with articulated deformations for pose estimation and retrieval [EB/OL]. 2021: *arXiv*: 2112.07334. https://arxiv.org/abs/2112.07334

[15] LIU L, DU J M, WU H, *et al*. Category-level articulated object 9D pose estimation *via* reinforcement learning[C]. *Proceedings of the* 31*st ACM International Conference on Multimedia. Ottawa ON Canada*. ACM, 2023: 728-736.

[16] JIANG H N, LIU L, OU L N, *et al*. EfficientCAPER: an end-to-end framework for fast and robust category-level articulated object pose estimation[C]. *Advances in Neural Information Processing Systems* 37. *December* 10-15, 2024. *Vancouver*, *BC*, *Canada*. *Neural Information Processing Systems Foundation*, *Inc*. (NeurIPS), 2024: 31968-31989.

[17] HE K M, ZHANG X Y, REN S Q, *et al*. Deep residual learning for image recognition[C]. 2016

IEEE Conference on Computer Vision and Pattern Recognition (*CVPR*). *June* 27-30, 2016, *Las Vegas*, *NV*, *USA*. IEEE, 2016: 770-778.

[18] QI C R, YI L, SU H, *et al*. PointNet++: deep hierarchical feature learning on point sets in a metric space[EB/OL]. 2017: *arXiv*: 1706.02413. https://arxiv.org/abs/1706.02413

[19] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. *June* 18-23, 2018, *Salt Lake City*, *UT*, *USA*. IEEE, 2018: 7132-7141.

[20] ZHENG L F, WANG C, SUN Y H, *et al*. HS-pose: hybrid scope feature extraction for category-level object pose estimation[C]. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 17-24, 2023, *Vancouver*, *BC*, *Canada*. IEEE, 2023: 17163-17173.

[21] CHARLES R Q, HAO S, MO K C, *et al*. PointNet: deep learning on point sets for 3D classification and segmentation[C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). *July* 21-26, 2017, *Honolulu*, *HI*, *USA*. IEEE, 2017: 77-85.

[22] XIANG F B, QIN Y Z, MO K C, *et al*. SAPIEN: a SimulAted part-based interactive ENvironment[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 13-19, 2020, *Seattle*, *WA*, *USA*. IEEE, 2020: 11094-11104.

作者简介：

欧林林(1980—),女,安徽宿州人,教授,博士生导师,2006年于上海交通大学获得博士学位,主要从事智能机器人系统、多机协同控制等前沿研究。E-mail: linlinou@zjut.edu.cn

陈　婷(2001—),女,浙江金华人,硕士研究生,2023年于浙江工业大学获得学士学位,主要从事类别级铰接物体位姿估计方法的研究。E-mail: 806082763@qq.com