



基于关键点增强的类别级物体6D姿态估计研究

熊艳彬^{1,2,3}, 王磊^{1,2,3}, 刘速杰^{1,2,3}, 谭明科⁴, 林飞⁴, 程俊^{1,2,3*}

1. 中国科学院深圳先进技术研究院, 深圳 518055

2. 中国科学院大学, 北京 101408

3. 香港中文大学, 香港 999077

4. 华坪隆基硅材料有限公司, 丽江 674800

* E-mail: jun.cheng@siat.ac.cn

收稿日期: 2025-11-19; 接受日期: 2026-02-11; 网络版发表日期: 2026-03-18

云南省科技计划项目(编号: 202302AD080008)、国家自然科学基金项目(批准号: U21A20487)、深圳市科技计划项目(编号: GJHZ20240218112504008)、深圳市医学研究专项(编号: C2401024)、广东省科技计划项目(编号: 2023TX07Z126)、深圳市高新区发展专项计划创新平台建设项目(高精度高分辨率 4D 成像概念验证中心)和中国科学院关键技术人才计划资助

摘要 类别级物体6维(6 Dimensions, 6D)姿态估计旨在预测来自类别已知的未见过目标实例的6D位姿和尺寸数据, 这要求模型对各种形状的物体实例具有强大的类内泛化能力. 尽管许多先前的方法都试图提高学习类内变化较大目标特征的能力, 但它们过于强调逐点稠密特征的学习或逐点坐标预测, 而缺少对目标整体几何特征和结构的学习, 导致此类方法对具有显著形状变化的未见过目标实例以及存在遮挡场景姿态估计的泛化性能较差. 为了解决此问题, 我们提出了一种基于关键点增强的类别级物体6D姿态估计方法. 该方法首先通过双模态分离跨模态注意力机制以及自适应权重融合的方式, 构建实例自适应关键点检测模块, 来学习一组稀疏关键点, 以表示实例的整体概略结构信息; 其次通过自适应动态邻域构建、线性和正弦位置编码以及多头注意力机制, 并通过条件归一化整合全局特征的方式, 构建感知特征聚合模块, 从而自适应地学习当前目标实例关键点的局部和全局特征. 关键点学习模块可以充分考虑和挖掘不同模态的信息, 以有效学习未见过实例的关键结构和概略信息. 特征聚合模块在关键点的基础上动态自适应地构建关键点局部和全局几何信息, 从而学习未见实例的几何信息, 提高对未见实例几何和结构信息的学习能力, 以及对未见目标实例姿态估计的泛化能力. 在REAL275和CAMERA25数据集上大量的实验结果表明, 所提出的关键点增强方法性能优异.

关键词 类别级物体6D姿态估计, 关键点增强, 跨模态注意力机制, 自适应权重融合, 自适应动态邻域, 正弦位置编码, 条件归一化

1 引言

物体6维(6 Dimensions, 6D)姿态估计旨在从输入图像(RGB/RGB-D)中预测物体的旋转、平移和大

小^[1~5]. 这项任务因其基础性以及在机器人操作和感知^[6~8]、AR/VR系统^[9,10]以及整体场景理解^[11,12]等领域中的巨大潜力而备受关注. 为了实现物体的姿态估计, 大量的科研探索在实例级物体6D姿态估计^[13~17]上取

引用格式: 熊艳彬, 王磊, 刘速杰, 等. 基于关键点增强的类别级物体6D姿态估计研究. 中国科学: 技术科学, 2026, 56: 721–736

Xiong Y B, Wang L, Liu S J, et al. 6D pose estimation of category-level objects based on keypoint enhancement (in Chinese). Sci Sin Tech, 2026, 56: 721–736, doi: 10.1360/SST-2025-0356

得了显著进展, 但对目标实例CAD模型的依赖阻碍了它们对新目标实例的推广. 为了解决这个问题, 标准化对象坐标空间方法(normalized object coordinate space, NOCS)^[1]中引入了类别级物体6D姿态估计, 其核心方法是在不使用当前目标CAD模型的情况下估计特定类别内未见过的物体的姿态. 本文中的方法将在基于关键点两阶段类别级物体6D姿态估计方法进行探索.

1.1 实例级物体姿态估计

实例级物体6D姿态估计侧重于根据已知对象的CAD模型确定其3D旋转和3D平移. 该领域的现有方法大致可分为三类: 基于直接姿态回归的方法^[18,19]、基于直接投票的方法^[20,21]和通过关键点检测或像素级3D坐标估计建立2D-3D对应的方法^[22,23]. 直接姿态回归方法利用端到端网络从提取的视觉特征中直接预测对象的旋转和平移参数. 直接投票方法在像素或点级别估计候选姿势及其相应的置信度分数, 最终选择具有最高置信度的姿势. 第三类实例级方法采用逐点或关键点对应的方法, 通常在对象CAD模型上预先定义一组关键点, 并检测它们在输入图像中的相应位置, 从而在CAD模型和输入图像之间建立密集的2D-3D对应关系, 然后使用透视图(perspective-n-point, PnP)算法或最小二乘优化等算法来估计对象姿态. 其中, 基于关键点对应的方法与我们的方法最相关. 然而, 实例级方法需要依赖精确的CAD模型, 而我们的方法旨在自适应地检测不同实例的关键点, 因为类别级物体6D姿态估计任务的推理过程中无法访问CAD模型.

1.2 类别级物体姿态估计

与依赖已知CAD模型的实例级方法不同, 类别级6D对象姿态估计旨在从预定义的对象类别中估计以前未见过的实例的姿态和大小. 由于对象实例之间存在大量的类内差异, 此设置本质上更加通用且更具挑战性. 当前阶段类别级物体6D姿态估计可以大致分为基于直接回归方法和基于点对应的方法.

1.2.1 基于直接回归的方法

直接回归方法采用单阶段方式, 旨在从输入对象观测中提取位姿敏感的特征, 然后直接预测对象姿态. FS-Net^[24]采用3D图卷积(3D graph convolution, 3DGC)

自动编码器进行方向特征提取, 并将旋转解耦为两个正交向量以处理对称对象. GPV-Pose^[25]和HS-Pose^[26]分别通过关注几何一致性和几何结构感知来增强姿势敏感特征提取. VI-Net^[27]通过将旋转解耦为视点和平面内分量来简化球体上的任务, 而SecondPose^[3]通过利用SE(3)一致的语义和几何特征提取进一步改进它. 直接回归方法不依赖于坐标映射过程, 使得他们可以端到端训练, 推理速度非常快. 但是当处理未见目标以及类内形变较大的物体时, 直接回归方法面临较大的挑战.

1.2.2 基于点对应的方法

为了缓解基于直接回归类方法的问题, 基于点对应的方法首先通过显式或隐式方式建立物体在相机坐标系和世界坐标系之间稠密点对应关系, 然后通过姿态拟合算法计算物体姿态. NOCS^[1]首先引入标准化对象坐标空间作为共享规范表示, 以对齐给定类别中的不同对象实例. SPD^[28]通过变形预先学习的类别形状先验来重建当前输入实例的3D模型, 然后预测观察到的实例点云与重建的对象点云之间的对应关系. GSNet^[29]通过当前实例多模态特征引导类别先验, 显式建立输入与输出空间的逐点对应关系. 后续相关工作^[30~32]进一步通过强化特征提取和形状先验的变形, 提高了基于先验的方法的性能和泛化能力.

然而, 该形状先验是在NOCS中针对每个类别预先学习的一组点云. 该点云是特定类别中对象的密集表示, 这不仅是冗余的, 而且使得网络计算成本昂贵. 此外, 形状先验在模型训练期间保持静态. 为了解决此问题, Query6DoF^[4]引入了与姿态估计网络联合优化动态可学习查询, 使形状先验能够在训练期间动态自适应. 此外, 无形形状先验方法^[33~35]因其改进的对各种不可见的物体几何形状的泛化而引起了越来越多的关注. IST-Net^[33]采用隐式空间映射方法, 直接学习相机空间和世界空间特征之间的对应关系, 而不依赖于显式先验. CLIPose^[34]利用视觉语言模型从文本和图像中提取丰富的语义特征, 通过对比学习将它们与点云表示对齐. SpherePose^[35]利用球面HEALPix^[36]网格作为代理形状, 学习观测点与球面NOCS之间的形状无关对应关系, 来提升类别级物体6D位姿的泛化与精度. 然而, 当前阶段大部分基于点对应的方法主要集中在稠密的逐点预测, 当遇到类内大形变目标时, 模

型的泛化能力会受到较大影响. 为此, AG-Pose^[2]提出检测实例自适应关键点并建立关键点对应关系, 以增强针对大形变物体的鲁棒性. 虽然当前方法已经取得了很大的进步, 但是在不同模态数据特征提取以及动态特征构造方面仍存在问题, 导致模型泛化性不足.

1.3 本文主要研究内容

受上述讨论的启发, 我们在AG-Pose^[2]的基础上, 提出了一种增强的实例自适应和几何感知关键点学习方法, 用于类别级物体6D姿态估计. 该方法在两个核心模块上进行了关键改进: 增强的实例自适应关键点检测模块和动态几何感知特征聚合模块.

增强的关键点检测模块旨在更有效地自适应检测各种实例的关键点. 与直接拼接点云和RGB特征进行单一注意力机制不同, 我们提出了双模态分离的跨模态注意力机制. 具体来说, 我们首先将点云特征和RGB特征分别投影到统一的特征空间, 然后初始化一组类别共享的可学习查询, 并将这些查询复制为两份, 分别与点云和RGB特征进行独立的跨模态注意力交

互. 每个分支都能学习到针对各自模态的实例自适应检测器, 从而充分挖掘不同模态数据特征. 为了自适应地平衡点云和RGB模态的贡献, 从而避免出现过度依赖某个模态信息的问题, 我们进一步设计了自适应模态门控单元, 根据当前特征动态生成模态权重, 并使用门控权重对两个分支的查询结果进行加权融合, 得到最终的实例自适应检测器. 随后, 通过计算实例自适应检测器和融合对象特征之间的余弦相似度来获得关键点的热图. 这种双模态分离的设计使得模型能够更好地利用不同模态的互补信息, 避免直接拼接导致的特征混淆, 从而提升关键点检测的质量和鲁棒性. 整体流程如图1所示. 此外, 我们保留了多样性损失和对对象感知的倒角距离损失来约束关键点的分布, 并引入了模态权重熵正则化损失和热图熵正则化损失, 进一步提高了关键点定位的稳定性和鲁棒性.

特征生成模块旨在更有效地提取检测到的关键点的局部和全局几何信息. 与原始方法中采用固定K近邻和基于余弦相似度的简单注意力机制不同, 我们提出了动态几何感知特征聚合模块. 具体来说, 为了合

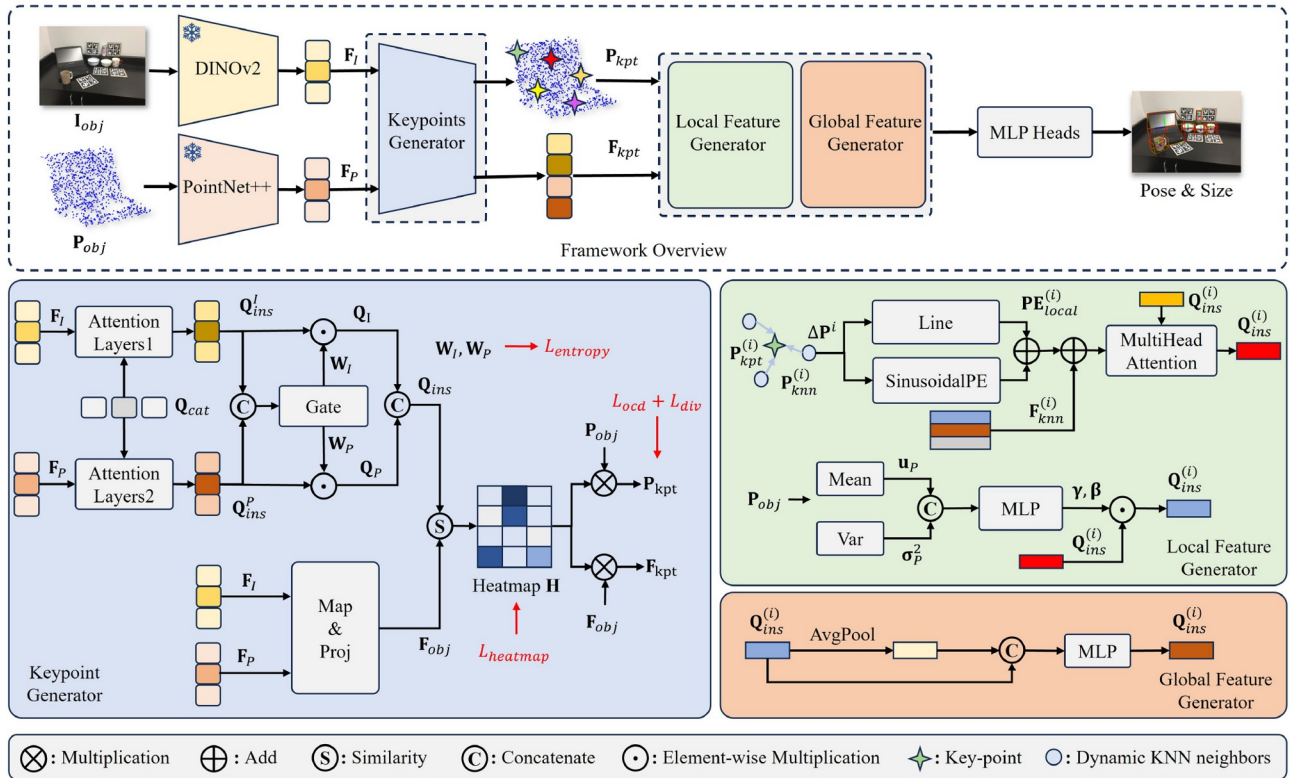


图 1 (网络版彩图) Framework Overview是我们提出方法的整体流程图

Figure 1 (Color online) Framework Overview is the overall overview of our proposed method.

并局部几何信息, 采用自适应邻域选择策略, 根据局部点云密度自适应地确定每个关键点的邻域大小. 我们首先计算每个关键点与所有点的距离, 基于距离分布的5%分位数确定自适应阈值, 并计算有效邻域密度, 然后将密度归一化后映射到预设范围, 得到动态邻域大小. 这种自适应策略使得在点云稀疏区域选择较大的邻域, 在密集区域选择较小的邻域, 从而更好地适应不同几何密度的场景. 为了编码局部几何关系, 我们提出了双重位置编码机制, 将多层感知机(multi-layer perceptron, MLP)位置编码和正弦位置编码相结合, 既保留了学习到的几何特征, 又通过正弦编码提供了稳定的高频几何细节表示. 在特征聚合阶段, 采用多头注意力机制替代传统的基于余弦相似度的简单注意力, 能够同时关注多个几何子空间, 捕获不同尺度的局部几何模式. 为了处理动态邻域大小, 我们构造键填充掩码来屏蔽无效邻居, 并通过残差连接和层归一化确保特征更新的稳定性. 为了将全局几何信息注入到关键点特征中, 我们采用特征线性调制机制, 根据点云的全局统计信息(均值和方差)自适应地调整关键点特征, 使得模型能够根据每个实例的整体形状和尺度自适应地调整特征, 提高对不同几何变化的鲁棒性. 最后, 我们整合所有关键点的全局信息, 实现局部几何聚合和全局特征融合的两阶段特征增强.

通过对两阶段模块的增强操作, 我们的方法可以有效地学习实例几何感知关键点, 以建立鲁棒的关键点级别对应关系. 关键点学习模块通过双模态分离注意力机制可以充分考虑和挖掘不同模态的信息, 以有效学习未见过实例的关键结构概略信息. 特征聚合模块在关键点的基础上动态自适应地构建关键点局部和全局几何信息, 从而学习未见实例的几何信息, 提高对未见实例姿态估计的泛化能力.

综上所述, 我们的创新点和贡献如下所述:

(1) 我们提出了增强的实例自适应和几何感知关键点学习方法, 通过双模态分离的跨模态注意力机制和自适应模态门控融合, 使模型可以充分感知来自不同模态的信息, 显著提升了关键点检测的质量和鲁棒性.

(2) 我们提出了动态几何感知特征聚合模块, 通过自适应邻域选择、双重位置编码、多头注意力机制、特征线性调制机制以及条件归一化, 有效地将局部和全局几何信息合并到关键点特征中, 建立了更加鲁棒的实例特征表达.

(3) 在标准数据集上的实验结果表明, 相比其他先进方法, 我们提出的方法在多个评估指标上均取得了有效提升, 验证了所提出方法的有效性.

2 关键点增强的类别级姿态估计算法设计

2.1 逐点特征提取

给定RGB-D图像, 类别级物体6D姿态估计的目标是预测给定类别物体的旋转 $\mathbf{R} \in SO(3)$, 平移 $\mathbf{t} \in \mathbb{R}^3$ 和大小 $\mathbf{s} \in \mathbb{R}^3$. 我们使用现成的语义分割方法Mask R-CNN^[37]从RGB图像中分割出目标对象, 然后结合目标对象的深度数据获取当前目标的3D点云, 最后将目标对象对应的RGB图像和3D点云作为模型的原始输入. 对于两种模态的输入数据, 我们采用逐点特征编码的方式提取对应像素的信息特征. 具体来说, 对于给定的目标RGB图像 $\mathbf{I}_{obj}$, 采用现有方法DINOv2^[38]提取图像的像素级语义和外观特征 $\mathbf{F}_I \in \mathbb{R}^{N \times C_1}$; 而对应的目标点云 $\mathbf{P}_{obj} \in \mathbb{R}^{N \times 3}$, 我们通过PointNet++^[39]从点云中获得几何特征 $\mathbf{F}_P \in \mathbb{R}^{N \times C_2}$. 进一步将图像和点云逐点特征对齐作为后续模块的输入.

2.2 基于跨模态交互的关键点生成

几何信息对于建立稳健的对应关系是必不可少的. 为每个点提供几何信息的直观想法通常采用普通注意力机制^[40]来聚合所有其他点的特征. 但是, 之前大部分方法构建稠密点之间的特征对应关系, 由于像素点数量较多, 导致计算成本较高; 此外, 当实例形状发生较大变化时, 稠密点对应的方式泛化性能将遇到较大挑战. 然而, 由于类别级姿态估计需要估计统一类别内不同形状的实例, 且在推理过程中无法访问实例模型, 无法采用固定关键点检测方法. 此外, 最远点采样方法无法进行端到端训练, 并且关键点检测经常会受到异常深度值、分割掩码噪声等因素影响, 而使关键点分布在异常点上. AG-Pose^[2]已经通过关键点生成和关键点几何特征构造的两阶段方式来缓解稠密点对应带来的问题, 但是在关键点学习阶段, 仍然存在不同模态信息挖掘不足、关键点几何特征构造不够灵活等问题. AMFNet^[41]通过自适应多特征融合的双重退化修复网络, 融合图像多尺度信息用于图像修复.

BGM-CGAN^[42]采用多模态融合的方式生成多样性噪声用于机械手充分感知外界环境. 为了解决以上问题, 我们采用自适应多模态思想, 设计了一个基于跨模态交互的实例自适应关键点检测模块来自适应地检测不同形状实例的关键点, 既可以充分利用不同模态是输入信息, 又可以灵活有效地表示对象整体形状, 并降低关注异常点的可能性.

2.2.1 双模态分离的跨模态注意力机制

具体来说, 首先初始化一组类别共享的可学习查询 $\mathbf{Q}_{cat} \in \mathbb{R}^{N_{kpt} \times C}$, 每个查询代表一个关键点检测器. 与原始方法直接拼接点云和RGB特征不同, 我们采用双模态分离的注意力机制(bimodal separation attention mechanism, BSAM), 分别处理点云和RGB数据, 以更好地利用不同模态的互补信息. 首先, 我们将通过基线模型提取的RGB特征 $\mathbf{F}_l \in \mathbb{R}^{N \times C_l}$ 和点云特征 $\mathbf{F}_p \in \mathbb{R}^{N \times C_2}$ 分别投影到统一的特征维度 d_{model} :

$$\mathbf{F}_l^{proj} = \mathbf{F}_l \mathbf{W}_l^{proj}, \quad (1)$$

$$\mathbf{F}_p^{proj} = \mathbf{F}_p \mathbf{W}_p^{proj}, \quad (2)$$

其中, $\mathbf{W}_l^{proj} \in \mathbb{R}^{C_l \times d_{model}}$, $\mathbf{W}_p^{proj} \in \mathbb{R}^{C_2 \times d_{model}}$, 是可学习的投影矩阵, 将不同维度的模态特征映射到统一的 d_{model} 维空间(通常 $d_{model}=256$).

为了充分挖掘不同模态中包含的信息, 我们通过注意力层将RGB和点云特征分别注入可学习的查询中. 将类别共享查询 $\mathbf{Q}_{cat}$ 复制为两份, 分别与点云和RGB特征进行独立的跨模态注意力交互. 该过程将类别共享检测器 $\mathbf{Q}_{cat}$ 转换为以对象特征为条件的实例自适应检测, 通过多层注意力块的迭代更新, 每个分支都能学习到针对各自模态的实例自适应检测器, 注意力机制转换公式如下:

$$\mathbf{Q}_l = \mathbf{Q}_{cat} \mathbf{W}^q, \mathbf{K}_l = \mathbf{F}_l^{proj} \mathbf{W}^k, \mathbf{V}_l = \mathbf{F}_l^{proj} \mathbf{W}^v, \quad (3)$$

$$\mathbf{Q}_{ins}^l = \text{Norm}(\mathbf{Q}_{cat} + \text{softmax}(\mathbf{Q}_l \mathbf{K}_l^T) \mathbf{V}_l), \quad (4)$$

$$\mathbf{Q}_p = \mathbf{Q}_{cat} \mathbf{W}^q, \mathbf{K}_p = \mathbf{F}_p^{proj} \mathbf{W}^k, \mathbf{V}_p = \mathbf{F}_p^{proj} \mathbf{W}^v, \quad (5)$$

$$\mathbf{Q}_{ins}^p = \text{Norm}(\mathbf{Q}_{cat} + \text{softmax}(\mathbf{Q}_p \mathbf{K}_p^T) \mathbf{V}_p), \quad (6)$$

其中, $\mathbf{Q}_{ins}^l$ 和 $\mathbf{Q}_{ins}^p$ 分别代表RGB和点云分支, $\mathbf{W}$ 代表注意力权重.

2.2.2 自适应模态门控加权融合

为了自适应地平衡点云和RGB模态的贡献, 我们设计了一个模态门控单元(modal gating unit), 根据全局特征动态生成模态权重:

$$\mathbf{F}_l = \text{AvgPool}(\mathbf{F}_l), \mathbf{F}_p = \text{AvgPool}(\mathbf{F}_p), \quad (7)$$

$$\mathbf{W}_l, \mathbf{W}_p = \text{Gate}(\text{Concat}[\mathbf{F}_l, \mathbf{F}_p]), \quad (8)$$

其中, $\text{Concat}[]$ 代表不同特征沿维度方向进拼接, $\text{Gate}()$ 是一个两层线性网络, 输出归一化的权重 $\mathbf{W}_l + \mathbf{W}_p = 1$. 然后, 我们使用门控权重对两个分支的查询结果进行加权融合:

$$\mathbf{Q}_{ins} = \text{Fusion}(\text{Concat}[\mathbf{W}_l \cdot \mathbf{Q}_{ins}^l, \mathbf{W}_p \cdot \mathbf{Q}_{ins}^p]), \quad (9)$$

其中, $\text{Fusion}()$ 是一个线性投影层, 将拼接后的特征映射回 d_{model} 维, 得到最终的实例自适应检测器 $\mathbf{Q}_{ins} \in \mathbb{R}^{N_{kpt} \times d_{model}}$.

在双分支注意力各自完成跨模态对齐之后, 我们加入模态门控单元, 因为单纯拼接/相加会把两个模态在语义上硬绑定, 一旦某一模态噪声偏大或信息缺失, 会影响另一个模态的信息表达. 模型门控单元基于全局平均池化得到的 $\mathbf{F}_l, \mathbf{F}_p$ 捕获了当前实例在两种模态下的整体质量与能量分布, 再通过两层MLP输出 $\mathbf{W}_l$ 和 $\mathbf{W}_p$, 相当于给每个对象一次模态可信度评估. 如果场景里点云稀疏/遮挡严重, 门控逻辑会自动降低 $\mathbf{W}_p$, 让注意力更多依赖RGB支撑; 反之, 如果图像模糊或光照极差, $\mathbf{W}_l$ 会被抑制, 系统仍可依赖点云几何维持鲁棒性. 最后的 $\mathbf{Q}_{ins} = \text{Fusion}(\text{Concat}[\mathbf{W}_l \cdot \mathbf{Q}_{ins}^l, \mathbf{W}_p \cdot \mathbf{Q}_{ins}^p])$ 保留了两路查询的模态专长, 又通过线性投影把它们重新映射到统一的 d_{model} , 确保后续热图计算仍可复用余弦相似度公式. 这样一来, 关键点检测器不仅“看见”了两种模态, 还能在实例级别自适应地挑选更可靠的信息, 显著减少异常模态对关键点分布的干扰, 同时将两种模态的数据对齐^[43].

2.2.3 关键点热图生成

通过双模态分离的跨模态注意力机制获取特征之后, 我们构建融合对象特征 $\mathbf{F}_{obj}$, 并计算实例自适应检测器 $\mathbf{Q}_{ins}^l$ 和目标特征 $\mathbf{F}_{obj}$ 之间的余弦相似度, 以生成关键点热图:

$$\mathbf{F}_{obj} = \text{Proj}(\text{Concat}[\mathbf{F}_I^{proj}, \mathbf{F}_P^{proj}]), \quad (10)$$

$$\mathbf{H} = \text{softmax}\left(\frac{\mathbf{Q}_{ins} \mathbf{F}_{obj}}{\|\mathbf{Q}_{ins}\|_2 \|\mathbf{F}_{obj}\|_2^\gamma}\right), \quad (11)$$

其中, $\text{Proj}()$ 将拼接后的特征投影回 d_{model} , 参数 $\gamma = 0.1$ 用于控制分布的尖锐程度. 热图 $\mathbf{H} \in \mathbb{R}^{N_{kpt} \times N}$ 表示每个关键点对每个点云位置的注意力权重. 随后, 通过加权求和得到相机空间下关键点的3D坐标 $\mathbf{P}_{kpt} \in \mathbb{R}^{N_{kpt} \times 3}$ 及其对应特征 $\mathbf{F}_{kpt} \in \mathbb{R}^{N_{kpt} \times C}$,

$$\mathbf{P}_{kpt} = \mathbf{H} \times \mathbf{P}_{obj}, \quad (12)$$

$$\mathbf{F}_{kpt} = \mathbf{H} \times \mathbf{F}_{obj}. \quad (13)$$

2.2.4 关键点优化监督

我们的目标是利用一组稀疏关键点来表示物体的整体结构及几何信息, 因此需要检测到的关键点尽可能均匀地分布在物体的表面, 以及关键代表性位置. 然而, 如不对关键点进行监督, 当前关键点往往过度聚集在小区域中, 如法向量变化比较大的区域, 并且经常集中在非表面或异常点上.

为了鼓励关键点尽量分布在物体表面不同区域, 需要采用多样性损失 L_{div} 来强制检测到的关键点彼此远离^[2], 公式如下:

$$L_{div} = \sum_{i=1}^{N_{kpt}} \sum_{j \neq i}^{N_{kpt}} d(\mathbf{P}_{kpt}^{(i)}, \mathbf{P}_{kpt}^{(j)}), \quad (14)$$

$$d(\mathbf{P}_{kpt}^{(i)}, \mathbf{P}_{kpt}^{(j)}) = \max(th_1 - \|\mathbf{P}_{kpt}^{(i)} - \mathbf{P}_{kpt}^{(j)}\|_2, 0), \quad (15)$$

其中, th_1 是超参数(通常设为0.01), 代表 $\mathbf{P}_{kpt}^{(i)}$ 第 i 个关键点. 该损失函数惩罚距离小于阈值的关键点对, 鼓励关键点在空间中分散分布. 为了鼓励关键点定位在物体表面并同时排除异常值, 我们仍然采用AG-Pose^[2]的设计, 通过物体感知的倒角距离损失 L_{ocd} 来约束 $\mathbf{P}_{kpt}$ 的分布. 形式上首先利用地面真值姿态和大小 $\mathbf{R}_{gt}$, $\mathbf{t}_{gt}$, $\mathbf{s}_{gt}$ 将 $\mathbf{P}_{obj}$ 转换到NOCS空间中, 并基于实例模型去除离群点以产生 $\mathbf{P}_{obj}^*$.

$$\mathbf{P}_{obj}^* = \left\{ \mathbf{x}_i \mid \mathbf{x}_i \in \mathbf{P}_{obj} \text{ and } \min_{\mathbf{y}_i \in M_{obj}} \left\| \frac{1}{\|\mathbf{s}_{gt}\|_2} \mathbf{R}_{gt}(\mathbf{x}_i - \mathbf{t}_{gt}) - \mathbf{y}_i \right\|_2 < th_2 \right\}. \quad (16)$$

然后, 我们通过倒角距离约束关键点接近 $\mathbf{P}_{obj}^*$, 使得关键点检测模块可以在推理过程中自动学习过滤掉异常点:

$$L_{ocd} = \text{Chamfer}(\mathbf{P}_{kpt}, \mathbf{P}_{obj}^*), \quad (17)$$

其中, $\text{Chamfer}()$ 代表倒角距离公式.

为了进一步防止模态门控单元生成的权重 $\mathbf{W}_I$ 和 $\mathbf{W}_P$ 极端化(如 $\mathbf{W}_I \rightarrow \mathbf{1}$, $\mathbf{W}_P \rightarrow \mathbf{0}$ 或相反), 即避免模型过度依赖图像或者点云信息, 我们引入模态权重熵正则化损失:

$$L_{entropy} = -\frac{1}{B} \sum_{b=1}^B [\mathbf{W}_I^b \log(\mathbf{W}_I^b + \varepsilon) + \mathbf{W}_P^b \log(\mathbf{W}_P^b + \varepsilon)], \quad (18)$$

其中, $\varepsilon = 10^{-7}$ 是数值稳定性常数. 该损失鼓励权重分布更均匀, 避免模型过度依赖单一模态, 提高在单模态缺失时的鲁棒性. 为了防止关键点热图 $\mathbf{H}$ 过度稀疏(接近one-hot分布)或过度平滑(接近均匀分布), 我们引入热图熵正则化损失:

$$L_{heatmap} = -\frac{1}{BN_{kpt}N} \times \sum_{b=1}^B \sum_{i=1}^{N_{kpt}} \sum_{j=1}^N [\mathbf{H}_{b,i,j} \log(\mathbf{H}_{b,i,j} + \varepsilon) + (1 - \mathbf{H}_{b,i,j}) \log((1 - \mathbf{H}_{b,i,j}) + \varepsilon)]. \quad (19)$$

该损失保持热图的适度集中性, 既不过于稀疏(导致关键点定位过于集中), 也不过于平滑(导致关键点定位不明确), 提高关键点定位的稳定性和鲁棒性.

2.3 动态几何感知特征聚合

检测到每个对象的关键之后, 可以直接预测这些关键点在NOCS空间中的坐标, 以建立关键点点级别的对应关系. 但是关键点离散的分布在物体不同位置, 它们之间是没有直接联系的孤立点, 也不代表当前实例的几何特征. 这可能导致在处理未见过的实例以及部件被遮挡的实例时存在大量错误对应. 因此, 为了更好地表达当前实例特征, 需要在关键点的基础上构建关键点局部和全局几何特征. 为此, 我们在AG-Pose^[2]的基础上进一步提出了一个动态几何感知特征聚合(dynamic geometric aware feature aggregation, DGAFA)模块, 以有效地获取关键点特征的几何信息合, 从而显式地构建当前实例的几何结构. 我们采用

两阶段方式进行几何特征聚合, 如图1所示.

2.3.1 自适应邻域构建与位置编码

本质上, 关键点局部几何信息可以用关键点与其相邻点之间的相对位置来表示, 而全局几何信息可以用关键点之间的相对位置来表示. 因此为了构建关键点局部几何特征, 我们首先在输入点云 $\mathbf{P}_{obj}$ 中为每个关键点选择最近的 K 个近邻 $\mathbf{P}_{knn} \in \mathbb{R}^{N_{kpt} \times k \times 3}$ 及其对应的特征 $\mathbf{F}_{knn} \in \mathbb{R}^{N_{kpt} \times k \times C}$. 但是不同实例的几何特征各异, 尤其是当实例形状变化较大时. 此外不同关键点周围的几何特性和与邻域的关系也不同, 如平滑区域周围邻域点较少, 陡峭区域的邻域点较多.

为了更好地表达关键点局部特征, 与传统的固定 K 近邻方法不同, DGAF采用动态邻域选择策略, 根据局部点云密度自适应地确定每个关键点的邻域大小. 具体而言, 对于第 i 个关键点 $\mathbf{P}_{kpt}^{(i)}$, 我们首先计算其与所有点 $\mathbf{P}_{obj}$ 的距离矩阵:

$$\mathbf{D}^{(i)} = \|\mathbf{P}_{kpt}^{(i)} - \mathbf{P}_{obj}\|_2 \in \mathbb{R}^N. \quad (20)$$

然后, 我们基于距离分布的分位数(如5%分位数)确定自适应阈值 $\gamma^{(i)}$, 并计算有效邻域密度:

$$\gamma^{(i)} = \text{Quantile}(\mathbf{D}^{(i)}, 0.05), \quad (21)$$

$$\rho^{(i)} = \sum_{j=1}^N \mathbf{1}[\mathbf{D}_j^{(i)} < \gamma^{(i)}], \quad (22)$$

其中, $\text{Quantile}()$ 计算分位数. 动态邻域大小 $k_{dyn}^{(i)}$ 根据密度归一化后映射到预设范围 $[k_{mi}, k_{ma}]$:

$$k_{dyn}^{(i)} = \text{Clip}\left(\frac{\rho^{(i)}}{\max \rho^{(i)}}(k_{ma} - k_{mi}) + k_{mi}, k_{mi}, k_{ma}\right), \quad (23)$$

其中, k_{mi} 和 k_{ma} 是预设的最小和最大邻域大小(如 $k_{mi} = 8$, $k_{ma} = 32$), $\text{Clip}()$ 为上下限裁剪函数. 这种自适应策略使得在点云稀疏区域选择较大的邻域, 在密集区域选择较小的邻域, 从而更好地适应不同几何密度的场景. 基于动态邻域大小, 我们选择 $k_{dyn}^{(i)}$ 个最近邻点及其对应特征:

$$\mathbf{P}_{knn}^{(i)} \in \mathbb{R}^{k_{dyn}^{(i)} \times 3}, \quad (24)$$

$$\mathbf{F}_{knn}^{(i)} \in \mathbb{R}^{k_{dyn}^{(i)} \times d_{model}}. \quad (25)$$

此外, 为了编码局部几何关系, 我们计算关键点与邻域点之间的相对位置, 并通过线性和正弦双重位置编码机制进行编码:

$$\Delta \mathbf{P}^{(i)} = \mathbf{P}_{kpt}^{(i)} - \mathbf{P}_{knn}^{(i)} \in \mathbb{R}^{k_{dyn}^{(i)} \times 3}, \quad (26)$$

$$\mathbf{PE}_{MLP}^{(i)} = \text{Line}(\Delta \mathbf{P}^{(i)}) \in \mathbb{R}^{k_{dyn}^{(i)} \times d_{model}}, \quad (27)$$

$$\mathbf{PE}_{sin}^{(i)} = \text{SinusoidalPE}(\Delta \mathbf{P}^{(i)}) \in \mathbb{R}^{k_{dyn}^{(i)} \times d_{model}}, \quad (28)$$

其中, $\text{Line}()$ 代表线性变换, $\text{SinusoidalPE}()$ 代表正弦位置编码. 最终的局部位置编码为两者的叠加:

$$\mathbf{PE}_{local}^{(i)} = \mathbf{PE}_{MLP}^{(i)} + \mathbf{PE}_{sin}^{(i)}. \quad (29)$$

这种双重编码机制既保留了MLP学习到的任务特定几何模式, 又通过正弦编码提供了稳定的高频几何细节表示.

2.3.2 基于多头注意力的局部几何聚合

在获得邻域特征和位置编码后, 我们采用多头注意力机制进行局部几何特征聚合, 替代传统的基于余弦相似度的简单注意力. 首先, 我们将关键点特征 $\mathbf{Q}_{ins}^{(i)}$ 作为查询(Query), 将增强后的邻域特征 $\mathbf{F}_{knn}^{en} = \mathbf{F}_{knn}^{(i)} + \mathbf{PE}_{local}^{(i)}$ 作为键(key)和值(value):

$$\tilde{\mathbf{Q}}_{ins}^{(i)} = \text{MultiHeadAttn}(\mathbf{Q}_{ins}^{(i)}, \mathbf{F}_{knn}^{en}, \mathbf{F}_{knn}^{en}), \quad (30)$$

$$\mathbf{Q}_{ins}^{(i)} = \text{MultiHeadAttn}(\tilde{\mathbf{Q}}_{ins}^{(i)} + \mathbf{Q}_{ins}^{(i)}), \quad (31)$$

其中, $\text{MultiHeadAttn}()$ 为多头注意力机制, 能够同时关注多个几何子空间, 捕获不同尺度的局部几何模式. 通过残差连接和层归一化, 我们确保特征更新的稳定性和梯度流动的顺畅性.

2.3.3 条件归一化与全局几何注入

为了将全局几何信息注入到关键点特征中, 我们采用特征线性调制(feature-wise linear modulation, FiLM), 根据点云的全局统计信息自适应地调整关键点特征. 具体而言, 对于当前批次中的每个实例, 我们首先计算点云 $\mathbf{P}_{obj} \in \mathbb{R}^{N \times 3}$ 的全局统计量:

$$\mathbf{u}_p = \text{Mean}(\mathbf{P}_{obj}) \in \mathbb{R}^3, \sigma_p^2 = \text{Var}(\mathbf{P}_{obj}) \in \mathbb{R}^3, \quad (32)$$

$$\mathbf{s}_{global} = \text{Concat}(\mathbf{u}_p, \sigma_p^2) \in \mathbb{R}^6, \quad (33)$$

其中, $\text{Mean}()$ 和 $\text{Var}()$ 分别代表均值和方差计算. 然后,

通过MLP生成条件归一化参数:

$$[\gamma, \beta] = \text{MLP}(\mathbf{s}_{\text{global}}) \in \mathbb{R}^{d_{\text{model}}}. \quad (34)$$

最后, 应用FiLM调制:

$$\mathbf{Q}_{\text{ins}}^{(i)} = \mathbf{Q}_{\text{ins}}^{(i)} \odot (\mathbf{1} + \gamma) + \beta, \quad (35)$$

其中, $\odot$ 表示逐元素相乘, 这种条件归一化机制使得模型能够根据每个实例的整体形状和尺度自适应地调整关键点特征, 提高对不同几何变化的鲁棒性.

2.3.4 全局特征融合

关键点局部几何信息代表了实例局部几何关系和特性, 但是仅局部几何信息不足以处理不同实例之间的显着形状变化. 为了更好地表达当前实例形状, 还需要构建当前实例的全局几何结构信息. 为此我们计算关键点特征的全局平均池化:

$$\mathbf{Q}_{\text{ins}}^{\text{global}} = \text{AvgPool}(\mathbf{Q}_{\text{ins}}^{(i)}). \quad (36)$$

然后, 我们将局部特征、全局特征和条件归一化后的特征进行融合:

$$\mathbf{Q}_{\text{ins}}^{(i)} = \text{ReLU}(\text{MLP}(\text{Concat}[\mathbf{Q}_{\text{ins}}^{(i)}, \mathbf{Q}_{\text{ins}}^{\text{global}}]) + \mathbf{Q}_{\text{ins}}^{(i)}). \quad (37)$$

上述两阶段聚合允许关键点自适应地聚合来自邻居的局部几何特征和来自其他关键点的全局几何信息. 值得注意的是, 由于关键点的数量很少, 因此上述过程的计算效率很高.

3 模型目标优化及目标姿态估计

在获得几何感知关键点特征后, 按照DPDN^[31]中的方式, 我们使用MLP从 $\mathbf{Q}_{\text{ins}}$ 中预测关键点 $\mathbf{P}_{\text{kpt}}^{\text{nocs}} \in \mathbb{R}^{N_{\text{kpt}} \times 3}$ 的NOCS坐标, 并通过这组关键点对应关系回归最终的姿态和大小.

$$\mathbf{P}_{\text{kpt}}^{\text{nocs}} = \text{MLP}(\mathbf{Q}_{\text{ins}}), \quad (38)$$

$$\mathbf{f}_{\text{pose}} = \text{Concat}[\mathbf{P}_{\text{kpt}}, \mathbf{F}_{\text{kpt}}, \mathbf{P}_{\text{kpt}}^{\text{nocs}}, \mathbf{Q}_{\text{ins}}], \quad (39)$$

$$\mathbf{R}, \mathbf{t}, \mathbf{s} = \text{MLP}_R(\mathbf{f}_{\text{pose}}), \text{MLP}_t(\mathbf{f}_{\text{pose}}), \text{MLP}_s(\mathbf{f}_{\text{pose}}). \quad (40)$$

更多细节请参考DPDN^[31]. 我们使用6D向量^[44]来表示旋转, 按照HS-Pose^[26]中的方式来预测真值和点云平均值之间的残差平移. 整体损失函数如下:

$$L_{\text{all}} = \lambda_1 L_{\text{ocd}} + \lambda_2 L_{\text{div}} + \lambda_3 L_{\text{nocs}} + \lambda_4 L_{\text{pose}} + \lambda_5 L_{\text{entropy}} + \lambda_6 L_{\text{heatmap}}, \quad (41)$$

其中, $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6$ 是超参数, 用于平衡每一项的贡献. 对于 L_{pose} , 我们采用使用 L_1 损失, $L_{\text{pose}} = \|\mathbf{R}_{\text{gt}} - \mathbf{R}\|_2 + \|\mathbf{t}_{\text{gt}} - \mathbf{t}\|_2 + \|\mathbf{s}_{\text{gt}} - \mathbf{s}\|_2$. 我们通过使用真值 $\mathbf{R}_{\text{gt}}, \mathbf{t}_{\text{gt}}, \mathbf{s}_{\text{gt}}$ 将关键点 $\mathbf{P}_{\text{kpt}}^{\text{gt}}$ 在相机空间下的坐标投影到NOCS中来生成关键点 $\mathbf{P}_{\text{kpt}}$ 的NOCS坐标真值. 对于 L_{nocs} , 我们使平滑 L_1 损失 $\text{SmoothL}_1()$, 公式如下:

$$\mathbf{P}_{\text{kpt}}^{\text{gt}} = \frac{1}{\|\mathbf{s}_{\text{gt}}\|_2} \mathbf{R}_{\text{gt}}(\mathbf{P}_{\text{kpt}} - \mathbf{t}_{\text{gt}}), \quad (42)$$

$$L_{\text{nocs}} = \text{SmoothL}_1(\mathbf{P}_{\text{kpt}}^{\text{gt}}, \mathbf{P}_{\text{kpt}}^{\text{nocs}}). \quad (43)$$

表1展示了所提出算法的整体过程, 其中MaskRCNN()用于通过Mask R-CNN^[37]获得对象分割结果, 包含目标掩码 $\{\mathbf{M}_n\}_{n=1}^{N_0}$ 、边界框 $\{\mathbf{B}_n\}_{n=1}^{N_0}$ 、类别信息 $\{\mathbf{c}_n\}_{n=1}^{N_0}$, 其中 N_0 为当前RGB图像中的目标个数. ImageCrop指在RGB图像中进行图像裁剪, MaskProjection旨在将对象掩模投影到深度图像中以获得对应的对象点云. DINOv2表示RGB图像特征提取器, PointNet++表示点云特征提取器. Gate表示门控机制, CA表示跨模态注意力机制, Concat代表不同特征沿维度方向进行拼接, SinusoidalPE表示正弦编码, NOCSPredictor和PoseEstimator分别表示NOCS预测器和姿态估计器.

4 实验结果与分析

4.1 数据集

我们在CAMERA25和REAL275数据集^[1]上训练并评估我们的方法. CAMERA25是一个由上下文感知混合现实方法生成的合成数据集, 包含6个物体类别, 共计30万张合成图像, 其中2.5万张用于评估. REAL275是一个更具挑战性的真实世界数据集, 包含杂乱的物体、遮挡和各种光照条件, 共计7050张真实图像, 其中2750张用于评估.

4.2 评价指标

我们使用3D交并比(3D IoU)的平均精度均值, 阈

表 1 基于关键点增强的类别级物体6D姿态估计方法整体流程**Table 1** Overall workflow of 6D pose estimation of category-level objects based on key points enhancement

基于关键点增强的类别级物体6D姿态估计方法整体流程

输入: RGB图像 $\mathbf{I}_{obj}$ 和深度图像 $\mathbf{D}$

输出: 物体6D姿态

第1阶段: 关键点获取

1 $\{\mathbf{M}_n\}_{n=1}^{N_0}, \{\mathbf{B}_n\}_{n=1}^{N_0}, \{\mathbf{c}_n\}_{n=1}^{N_0} = \text{MaskRCNN}(\mathbf{I}_{rgb}),$ 语义分割2 for $n=1, 2, \dots, N_0$ do3 $\mathbf{I}_{obj} = \text{ImageCrop}(\mathbf{B}_n, \mathbf{I}_{rgb}),$ 目标裁剪4 $\mathbf{P}_{obj} = \text{MaskProjection}(\mathbf{M}_n, \mathbf{D}),$ 深度图像中的掩模投影5 $\mathbf{F}_I = \text{DINOv2}(\mathbf{I}_{obj}),$ 目标图像特征提取6 $\mathbf{F}_P = \text{PointNet}++(\mathbf{P}_{obj}),$ 目标点云特征提取7 $\mathbf{W}_I, \mathbf{W}_P = \text{Gate}(\mathbf{F}_I, \mathbf{F}_P),$ 图像和点云权重获取8 $\mathbf{Q}_{ins}^I = \text{CA}(\mathbf{F}_I, \mathbf{Q}_{cat}^I), \mathbf{Q}_{ins}^P = \text{CA}(\mathbf{F}_P, \mathbf{Q}_{cat}^P),$ 双模态注意力9 $\mathbf{Q}_{ins} = \text{Concat}(\mathbf{W}_I \mathbf{Q}_{ins}^I, \mathbf{W}_P \mathbf{Q}_{ins}^P),$ 实例自适应获取10 通过式(10)~(13)获取热图 $\mathbf{H}$ 以及关键点 $\mathbf{P}_{kpt}$

第2阶段: 关键点几何特征聚合

11 根据距离分布动态获取关键点的最近邻点个数 $k_{dyn}^{(i)}$ 12 $\mathbf{P}_{knn}^{(i)}, \mathbf{F}_{knn}^{(i)},$ 获取第 i 个关键点的邻域点及特征13 $\mathbf{PE}_{MLP}^{(i)} = \text{Line}(\mathbf{P}_{kpt}^{(i)} - \mathbf{P}_{knn}^{(i)}),$ 相对位置线性编码14 $\mathbf{PE}_{sin}^{(i)} = \text{SinusoidalPE}(\mathbf{P}_{kpt}^{(i)} - \mathbf{P}_{knn}^{(i)}),$ 相对位置正弦编码15 $\mathbf{PE}_{local}^{(i)} = \mathbf{PE}_{MLP}^{(i)} + \mathbf{PE}_{sin}^{(i)},$ 关键点局部位置编码整合16 $\mathbf{F}_{knn}^{en} = \mathbf{F}_{knn}^{(i)} + \mathbf{F}_{local}^{(i)},$ 关键点局部邻域特征增强17 $\mathbf{Q}_{ins}^{(i)} = \text{CA}(\mathbf{Q}_{ins}^{(i)}, \mathbf{F}_{knn}^{en}),$ 关键点局部特征聚合18 $\mathbf{Q}_{ins}^{(i)} = \mathbf{Q}_{ins}^{(i)} \odot (1 + \gamma) + \beta,$ 关键点局部特征条件归一化19 $\mathbf{Q}_{ins}^{global} = \text{AvgPool}(\mathbf{Q}_{ins}^{(i)}),$ 关键点特征全局平均池化20 $\mathbf{Q}_{ins}^{(i)} = \text{ReLU}(\text{MLP}(\text{Concat}(\mathbf{Q}_{ins}^{(i)}, \mathbf{Q}_{ins}^{global}))) + \mathbf{Q}_{ins}^{(i)}.$

第3阶段: 姿态估计

21 $\mathbf{P}_{kpt}^{noes} = \text{NOCSPredictor}(\mathbf{Q}_{ins}),$ NOCS空间关键点预测22 $\{\mathbf{R}, \mathbf{t}, \mathbf{s}\}_n = \text{PoseEstimator}(\mathbf{P}_{kpt}, \mathbf{F}_{kpt}, \mathbf{P}_{kpt}^{noes}, \mathbf{Q}_{ins}),$

23 end for

24 return $\{\mathbf{R}, \mathbf{t}, \mathbf{s}\}_{n=1}^{N_0}.$

值分别为50%和75%, 来评估物体的姿态和尺寸, 重叠率大于阈值被认为是正确的. 我们还使用与真实值的偏差来评估估计的旋转和平移, 阈值分别为 5° , 10° , 2 cm和5 cm. 具体来说, 我们使用上述旋转和平移阈值的组合来评估姿态估计性能: $5^\circ 2\text{cm}$, $5^\circ 5\text{cm}$, $10^\circ 2\text{cm}$ 和 $10^\circ 5\text{cm}$.

4.3 推理细节

为了公平比较, 使用分割模型Mask R-CNN^[37]获

取分割掩码. 特征提取之前将图像裁剪并调整为 192×192 , 点云中的点数 N 设置为1024. 对于模型参数, 特征维度 $C_1 = C_2 = 128$, $C = 256$. 关键点数量设置为 $N_{kpt} = 96$, 对于损失函数中的超参数, 关键点优化中的两个阈值设置为 $th_1 = 0.01$ 和 $th_2 = 0.1$, 并且 $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6$ 分别为1.0, 5.0, 1.0, 0.3, 0.1, 0.01. 我们在合成数据集和真实数据集上训练模型, 以便在REAL275上进行评估; 而仅在合成数据集上训练模型, 以便在CAMERA25上进行评估. 我们使用随机平移 $\Delta \mathbf{t} \sim U(-0.02, 0.02)$ 和随机缩放 $\Delta \mathbf{s} \sim U(-0.02, 0.02)$ 进行数据增强, 并对每个轴应用随机旋转, 旋转角度从 $U(0, 20)$ 采样. 训练过程使用ADAM^[45]优化器训练我们的网络, 并采用 $2e^{-5} \sim 5e^{-4}$ 范围内的三角循环学习率调度^[46]. 所有实验都在一块RTX3090GPU上进行, 一个批次的数据量为40.

4.4 与当前其他先进方法的定量比较

将我们的方法与其他先进的类别级物体6D姿态估计方法在REAL275与CAMERA25数据集上进行对比. 表2展示了我们所提出的方法在REAL275数据集上与其他优秀方法的定量对比, 其中加粗部分为最优指标. 结果显示, 我们的方法在所有指标上都优于所采用的基线方法AG-Pose^[2], 并且在核心指标上优于当前其他优秀的算法. 我们的方法不需要使用形状先验, 模型可以根据当前实例动态调整. 具体来说, 在最严格的 $5^\circ 2\text{cm}$ 度量指标上, 我们的方法精度达到59.5%, 比基线模型的精度提高2.5个百分点, 在 $10^\circ 2\text{cm}$ 度量指标上比基准模型提高了3.4个百分点. 而且我们的方法在所有指标上都远超最优秀的基于形状先验的方法. 至于先验无关方法, 在最严格的 $5^\circ 2\text{cm}$ 度量指标上, 我们的方法比IST-Net^[33]、Query6DoF^[4]、SecondPose^[3]、AG-Pose^[2]、GCE-Pose^[5]和SpherePose^[35]等其他优秀的方法分别提高了12, 10.5, 3.3, 2.5, 2.5, 1.3个百分点. 需要注意的是Query6DoF^[4]的目标是学习一组稀疏查询作为隐式形状先验, 查询特征用于更新点特征以建立更好的密集对应关系. 与此不同的是, 我们的方法旨在使用稀疏关键点集对对象的整体概略关键信息以及几何信息进行显式建模, 以建立稳健的关键点级对应关系并进行姿态估计. 我们方法的优越性能表明了几何信息在类别级物体6D姿态估计中的重要性. 此外, 我们还在

CAMERA25数据集上验证了所提方法的优越性, 如表3所示, 其中加粗部分为最优指标, 我们的方法在各项核心指标上表现最优.

4.5 可视化结果

我们对所提出的方法进行了定性可视化分析, 如图2和图3所示. 具体来说, 我们在真实数据和合成数据的不同场景下做了相关验证, 并可视化模型的预测姿态.

我们在REAL275数据集上将我们的方法与基线方法AG-Pose^[2]进行了可视化对比, 如图4所示. 绿色和红色分别代表真实边界框和预测边界框, 浅蓝色边框

内容为存在差异的地方. 可视化结果表明我们的方法在姿态估计精度方面表现出色. 图5将我们的方法在NOCS空间中预测的关键点与基线方法AG-Pose^[2]的预测结果进行了对比. 为了突出差异, 关键点根据其NOCS误差进行颜色编码, 从绿色到红色颜色数值呈性映射, 当关键点为绿色时误差为0, 红色时误差为0.2. 如图5所示, AG-Pose^[2]产生的关键点存在较高的NOCS误差和不均匀的表面覆盖, 而我们的方法预测的关键点更均匀, 误差也相对较小.

此外, 我们在REAL275真实数据集和CAMERA25合成数据集上对我们所提出方法的平均精度(mean average precision, mAP)在不同阈值上进行了彻底的评

表 2 不同方法在REAL275数据集上的定量比较(单位: %)

Table 2 Quantitative comparison of different methods on the REAL275 dataset (unit: %)

方法	是否使用形状先验	IoU ₅₀	IoU ₇₅	5°2cm	5°5cm	10°2cm	10°5cm
NOCS ^[1]	×	78	30.1	7.2	10	13.8	25.2
GPV-Pose ^[25]	×	—	64.4	32	42.9	—	73.3
IST-Net ^[33]	×	82.5	76.6	47.5	53.4	72.1	80.5
Query6DoF ^[4]	×	82.5	76.1	49	58.9	68.7	83
SecondPose ^[3]	×	—	49.7	56.2	63.6	74.7	86.0
AG-Pose ^[2]	×	84.1	80.1	57	64.6	75.1	84.7
GCE-Pose ^[5]	×	84.1	79.8	57.0	65.1	75.6	86.3
SpherePose ^[35]	×	84.0	79	58.2	67.4	76.2	88.2
SPD ^[28]	√	77.3	53.2	19.3	21.4	43.2	54.1
SGPA ^[30]	√	80.1	61.9	35.9	39.6	61.3	70.7
DPDN ^[31]	√	83.4	76	46	50.7	70.4	78.4
Ours	×	84.1	80.4	59.5	64.8	78.5	85.9

表 3 不同方法在CAMERA25数据集上的定量比较(单位: %)

Table 3 Quantitative comparison of different methods on the CAMERA25 dataset (unit: %)

方法	5°2cm	5°5cm	10°2cm	10°5cm
NOCS ^[1]	32.3	40.9	48.2	64.4
DualPoseNet ^[48]	64.7	70.7	77.2	84.7
GPV-Pose ^[25]	72.1	79.1	—	89
Query6DoF ^[4]	78	83.1	83.9	90
AG-Pose ^[2]	79.5	83.7	87.1	92.6
SPD ^[28]	54.3	59	73.3	81.5
SAR-Net ^[47]	66.7	70.9	75.3	80.3
SGPA ^[30]	70.7	74.5	82.7	88.4
Ours	80.8	84.6	88.4	93.6



图 2 (网络版彩图)REAL275数据集6个真实场景中预测姿态的定性结果, 绿色边框为真实值, 红色边框为预测值

Figure 2 (Color online) Qualitative results of predicting poses in 6 real scenarios of the REAL275 dataset, with the red line representing the predicted value and the green line representing the true value.

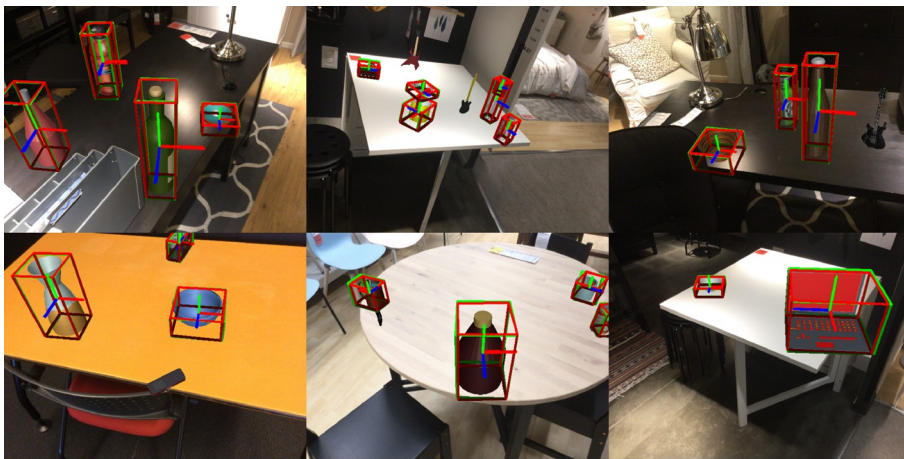


图 3 (网络版彩图)CAMERA25数据集不同场景中预测姿态的定性结果, 绿色边框为真实值, 红色边框为预测值

Figure 3 (Color online) Qualitative results of predicting poses in different scenes from the CAMERA25 dataset, with the red line representing the predicted value and the green line representing the true value.

估, 定量结果如图6所示. 结果表明, 我们的方法在所有评估阈值上始终如一地实现了卓越的性能, 在真实和合成环境中都表现出了特别强的度量稳定性.

4.6 消融实验

在本节中, 我们在REAL275真实数据集上进行消融实验与研究, 以证明所提出方法的有效性. 表4中BS代表双模态分离的注意力机制模块(BSAM), DG代表动态几何感知特征聚合模块(DGAFA), 其中加粗部分为最优指标. 从表4中可以看出, 只通过修改基于跨模

态交互的关键点生成模块, 在REAL275数据集上多数指标比基准模型有所提升. 单独修改动态几何感知特征聚合时, 在REAL275数据集上指标并没有提升太多, 因为如果不能提高关键点预测的准确性, 当预测的关键点出现偏差时, 通过自适应邻域构建的邻域特征和位置编码, 基于多头注意力的局部几何聚合等复杂操作会引入更多的噪声和误差, 从而导致性能不佳. 当跨模态交互的关键点生成模块和动态几何感知特征聚合模块同时修改时, 模型性能得到了较大幅度的提升, 说明更准确的关键点检测可以引导更好的特征提取过

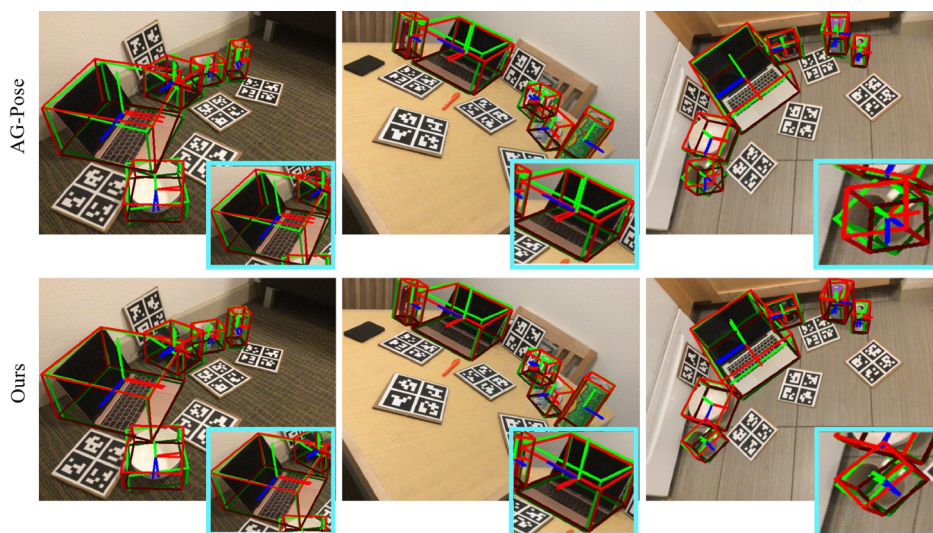


图 4 (网络版彩图)我们的方法与基线方法在REAL275数据集上的可视化对比, 第1行为AG-Pose^[2]的结果, 第2行为我们方法的结果. 可视化对比结果显示, 我们的方法表现优异

Figure 4 (Color online) Our method is compared with the baseline method on the REAL275 dataset. The first row shows the results of AG-Pose^[2], and the second row shows our results. The visualization demonstrates that our method outperforms the baseline method.

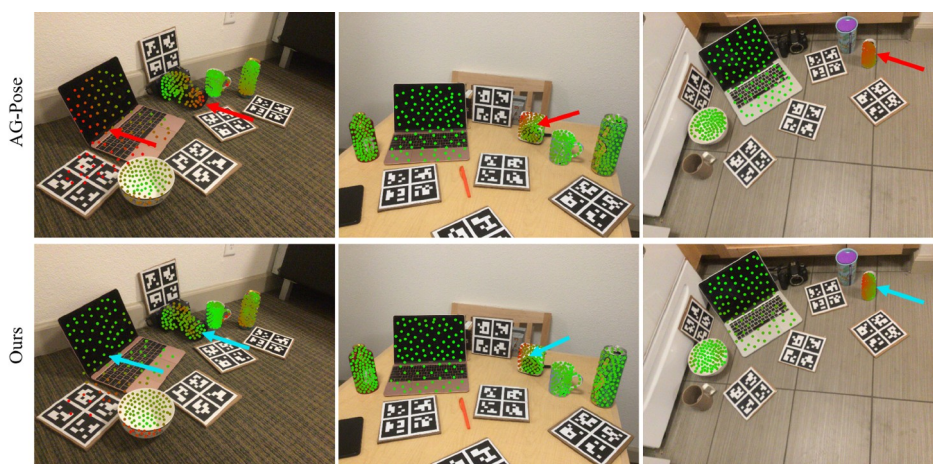


图 5 (网络版彩图)在REAL275数据集上将我们的方法与基线方法在NOCS空间中预测的关键点进行比较. 第1行显示基线方法的结果, 第2行显示我们方法的结果. 结果显示, 我们的方法预测的关键点更均匀且误差更小

Figure 5 (Color online) Our method compared with baseline methods in predicting key points in NOCS space on the REAL275 dataset. The first row shows the results of the baseline method, and the second row shows the results of our method. The results show that our method predicts key points more uniformly and with smaller errors.

程, 从而提高模型整体性能, 验证了我们所提出方法的有效性.

5 结论

我们在AG-Pose^[2]的基础上, 提出了一种增强的实例自适应和几何感知关键点学习方法, 用于类别级6D

对象姿态估计. 具体来说, 我们在关键点生成阶段提出了双模态分离的跨模态注意力机制, 通过独立的点云和RGB分支分别学习模态特异性表示, 并采用自适应模态门控单元实现更灵活的跨模态融合, 从而显著提升了关键点检测的质量和鲁棒性. 此外, 我们在关键点特征聚合阶段提出了动态几何感知特征聚合模块, 通过自适应邻域选择、多头注意力机制和条件归一

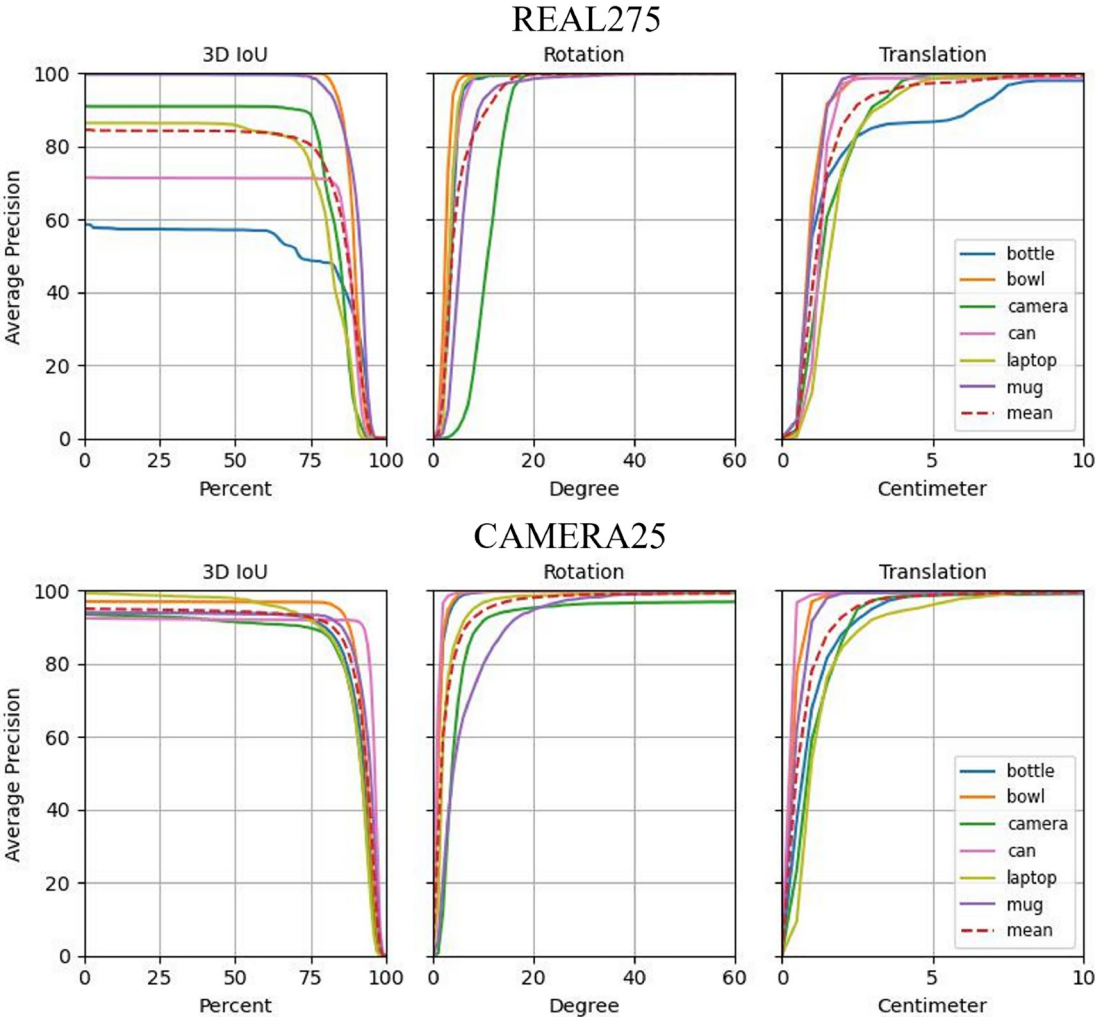


图 6 (网络版彩图)针对不同对象类别, 分别采用3D IoU、旋转阈值和平移阈值三种指标绘制平均精度均值曲线. 上图是 REAL275数据集结果, 下图是CAMERA25数据集结果

Figure 6 (Color online) For different object categories, mean accuracy curves were plotted using three metrics: 3D IoU, rotation threshold, and translation threshold. The top figure shows the results for the REAL275 dataset, and the bottom figure shows the results for the CAMERA25 dataset.

表 4 两个改进模块在REAL275数据集上的消融实验

Table 4 Ablation experiments of two improved modules on the REAL275 dataset

过程	BS	DG	5°2cm	5°5cm	10°2cm	10°5cm
Baseline	×	×	57	64.6	75.1	84.7
Stage-1	√	×	57.9	64.4	77.5	86.3
Stage-2	×	√	57.8	64.6	75.4	83.9
完整	√	√	59.5	64.8	78.5	85.9

化, 有效地将局部和全局几何信息合并到关键点特征中, 建立了更加鲁棒的关键点级对应关系. 我们在 REAL275和CAMERA25数据集上进行了全面的实验,

实验结果表明, 相比基线方法以及当前其他优秀的方法, 我们的方法在多个评估指标上均取得了提升, 验证了所提出方法的有效性.

参考文献

- 1 Wang H, Sridhar S, Huang J W, et al. Normalized object coordinate space for category-level 6D object pose and size estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019. 2642–2651
- 2 Lin X, Yang W F, Gao Y, et al. Instance-adaptive and geometric-aware keypoint learning for category-level 6D object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2024. 21040–21049
- 3 Chen Y M, Di Y, Zhai G Y, et al. SecondPose: SE(3)-consistent dual-stream feature fusion for category-level pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2024. 9959–9969
- 4 Wang R Q, Wang X G, Li T, et al. Query6DoF: Learning sparse queries as implicit shape prior for category-level 6DoF pose estimation. In: Proceedings of IEEE/CVF International Conference on Computer Vision. Paris, 2023. 14055–14064
- 5 Li W H, Xu H L, Huang J W, et al. GCE-Pose: Global context enhancement for category-level object pose estimation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville, 2025. 27154–27165
- 6 Zhai G Y, Cai X N, Huang D Y, et al. SG-Bot: Object rearrangement via coarse-to-fine robotic imagination on scene graphs. In: Proceedings of the IEEE International Conference on Robotics and Automation. Yokohama, 2024. 4303–4310
- 7 Wu C Z, Chen J, Cao Q Y, et al. Grasp proposal networks: An end-to-end solution for visual learning of robotic grasps. *Adv Neural Inf Process Syst*, 2020, 33: 13174–13184
- 8 Tong X B, Xiong P W, Chen Z Y, et al. Robotic tactile object perception based on adaptive multikernel sparse representation (in Chinese). *Sci Sin Tech*, 2023, 53: 2053–2065 [童小宝, 熊鹏文, 陈志远, 等. 基于自适应多核稀疏表示的机器人触觉物体感知研究. *中国科学: 技术科学*, 2023, 53: 2053–2065]
- 9 Marchand E, Uchiyama H, Spindler F. Pose estimation for augmented reality: A hands-on survey. *IEEE Trans Visual Comput Graphics*, 2015, 22: 2633–2651
- 10 Kim J, Oh H, Kim W, et al. A deep motion sickness predictor induced by visual stimuli in virtual reality. *IEEE Trans Neural Netw Learn Syst*, 2020, 33: 554–566
- 11 Di Y, Zhang C Y G, Zhang R D, et al. U-RED: Unsupervised 3D shape retrieval and deformation for partial point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, 2023. 8884–8895
- 12 Zhai G Y, Örnek E P, Wu S C, et al. CommonScenes: Generating commonsense 3D indoor scenes with scene graph diffusion. *Adv Neural Inf Process Syst*, 2023, 36: 30026–30038
- 13 He Y S, Sun W, Huang H B, et al. PVN3D: A deep point-wise 3D key points voting network for 6DoF pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020. 11632–11641
- 14 He Y S, Huang H B, Fan H, et al. FFB6D: A full flow bidirectional fusion network for 6D pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2021. 3003–3013
- 15 Peng S D, Liu Y, Huang Q X, et al. PVNet: Pixel-wise voting network for 6DoF pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019. 4561–4570
- 16 Wang G, Manhardt F, Tombari F, et al. GDR-Net: Geometry-guided direct regression network for monocular 6D object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2021. 16611–16621
- 17 Wu Y Z, Zand M, Etemad A, et al. Vote from the center: 6DoF pose estimation in RGB-D images by radial key point voting. In: Proceedings of the European Conference on Computer Vision. Tel Aviv, 2022. 335–352
- 18 Di Y, Manhardt F, Wang G, et al. SO-Pose: Exploiting self-occlusion for direct 6D pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, 2021. 12396–12405
- 19 Wang C, Xu D F, Zhu Y K, et al. DenseFusion: 6D object pose estimation by iterative dense fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019. 3343–3352
- 20 Tian M, Pan L, Ang M H, et al. Robust 6D object pose estimation by learning RGB-D features. In: Proceedings of the IEEE International Conference on Robotics and Automation. Paris, 2020. 6218–6224
- 21 Zhou G Y, Wang H Q, Chen J X, et al. PR-GCN: A deep graph convolutional network with point refinement for 6D pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, 2021. 2793–2802

- 22 Park K, Patten T, Vincze M. Pix2Pose: Pixel-wise coordinate regression of objects for 6D pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, 2019. 7668–7677
- 23 Xu Y, Lin K Y, Zhang G, et al. RNNPose: 6-DoF object pose estimation via recurrent correspondence field estimation and pose optimization. *IEEE Trans Pattern Anal Mach Intell*, 2024, 46: 4669–4683
- 24 Chen W, Jia X, Chang H J, et al. FS-Net: Fast shape-based network for category-level 6D object pose estimation with decoupled rotation mechanism. In: Proceedings of the IEEE/CVF Conference on Computer Cision and Pattern Recognition. Nashville, 2021. 1581–1590
- 25 Di Y, Zhang R D, Lou Z Q, et al. GPV-Pose: Category-level object pose estimation via geometry-guided point-wise voting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022. 6781–6791
- 26 Zheng L F, Wang C, Sun Y H, et al. HS-Pose: Hybrid scope feature extraction for category-level object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023. 17163–17173
- 27 Lin J H, Wei Z W, Zhang Y B, et al. VI-Net: Boosting category-level 6D object pose estimation via learning decoupled rotations on the spherical representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, 2023. 14001–14011
- 28 Tian M, Ang Jr. M H, Lee G H. Shape prior deformation for categorical 6D object pose and size estimation. In: Proceedings of the European Conference on Computer Vision. Glasgow, 2020. 530–546
- 29 Liu P L, Zhang Q S, Cheng J. GSNet: Model reconstruction network for category-level 6D object pose and size estimation. In: Proceedings of the IEEE International Conference on Robotics and Automation. London, 2023. 2898–2904
- 30 Chen K, Dou Q. SGPA: Structure-guided prior adaptation for category-level 6D object pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, 2021. 2773–2782
- 31 Lin J H, Wei Z W, Ding C X, et al. Category-level 6D object pose and size estimation using self-supervised deep prior deformation networks. In: Proceedings of the European Conference on Computer Vision. Tel Aviv, 2022. 19–34
- 32 Wang J Z, Chen K, Dou Q. Category-level 6D object pose estimation via cascaded relation and recurrent reconstruction networks. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. Prague, 2021. 4807–4814
- 33 Liu J H, Chen Y K, Ye X Q, et al. IST-Net: Prior-free category-level pose estimation with implicit space transformation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, 2023. 13978–13988
- 34 Lin X, Zhu M, Dang R, et al. CLIPose: Category-level object pose estimation with pre-trained vision-language knowledge. *IEEE Trans Circuits Syst Video Technol*, 2024, 34: 9125–9138
- 35 Ren H, Yang W, Liu X, et al. Learning shape-independent transformation via spherical representations for category-level object pose estimation. In: Proceedings of the International Conference on Learning Representations. Singapore, 2025
- 36 Gorski K M, Hivon E, Banday A J, et al. HEALPix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *Astrophys J*, 2005, 622: 759–771
- 37 He K M, Gkioxari G, Dollár P, et al. Mask R-CNN. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Venice, 2017. 2961–2969
- 38 Oquab M, Darcet T, Moutakanni T, et al. Dinov2: Learning robust visual features without supervision. arXiv: [2304.07193](https://arxiv.org/abs/2304.07193)
- 39 Qi C R, Yi L, Su H, et al. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In: Advances in Neural Information Processing Systems 30. 2017
- 40 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Advances in Neural Information Processing Systems 30. 2017
- 41 Zhao X Y, Peng C Y, Zhang X J. Dual image degradation repair network based on adaptive multifeature fusion (in Chinese). *Sci Sin Tech*, 2023, 53: 1939–1952 [赵雪雅, 彭春燕, 张效娟. 基于自适应多特征融合的双重图像退化修复网络. 中国科学: 技术科学, 2023, 53: 1939–1952]
- 42 Xiong P W, Tong X B, Song A G, et al. Robotic cross-modal generative adversarial network based on variational Bayesian Gaussian mixture noise model (in Chinese). *Sci Sin-Inf*, 2021, 51: 104–121 [熊鹏文, 童小宝, 宋爱国, 等. 基于变分贝叶斯高斯混合噪声模型的机器人跨模态生成对抗网络. 中国科学: 信息科学, 2021, 51: 104–121]
- 43 Dou M Y, Geng Y J, Yang J B. Alignment regression hand pose estimation network based on focused attention mechanism (in Chinese). *J Integr Technol*, 2025, 14: 64–77 [窦铭扬, 耿艳娟, 杨佳彬. 基于聚焦注意力机制的对齐回归手部姿态估计网络. 集成技术, 2025, 14: 64–77]
- 44 Zhou Y, Barnes C, Lu J W, et al. On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019. 5745–5753
- 45 Diederik P, Jimmy B. ADAM: A method for stochastic optimization. arXiv: [1412.6980](https://arxiv.org/abs/1412.6980)

- 46 Smith L N. Cyclical learning rates for training neural networks. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Santa Rosa, 2017. 464–472
- 47 Lin H T, Liu Z C, Cheang C, et al. SAR-Net: Shape alignment and recovery network for category-level 6D object pose and size estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022. 6707–6717
- 48 Lin J H, Wei Z W, Li Z H, et al. DualPoseNet: Category-level 6D object pose and size estimation using dual pose network with refined learning of pose consistency. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, 2021. 3560–3569

6D pose estimation of category-level objects based on keypoint enhancement

XIONG YanBin^{1,2,3}, WANG Lei^{1,2,3}, LIU SuJie^{1,2,3}, TAN MingKe⁴, LIN Fei⁴ & CHENG Jun^{1,2,3}

¹ Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

² University of Chinese Academy of Sciences, Beijing 101408, China

³ The Chinese University of Hong Kong, Hong Kong 999077, China

⁴ Huaping Longi Silicon Materials Co., Ltd., Lijiang 674800, China

Category-level object 6D pose estimation aims to predict 6D pose and size data for unseen target instances with known categories, requiring models to possess strong intra-category generalization capabilities for objects of diverse shapes. Although many previous methods have attempted to enhance the learning of features with significant intra-category variations, they overly emphasize point-wise dense feature learning or point-wise coordinate prediction while lacking the learning of target-wide geometric features and structures. This results in poor generalization performance for unseen target instances with significant shape variations and occluded scenarios in pose estimation. To address this issue, we propose a category-level object 6D pose estimation method based on keypoint enhancement. First, this method constructs an instance-adaptive keypoint detection module through a dual-modal separated cross-modal attention mechanism and adaptive weight fusion, learning a set of sparse keypoints to represent the overall schematic structural information of the instance. Second, it builds a perceptual feature aggregation module via adaptive dynamic neighborhood construction, linear and sinusoidal positional encoding, and multi-head attention mechanisms, integrating global features through conditional normalization to adaptively learning local and global features of the current target instance's keypoints. The keypoint learning module can fully consider and exploit information from different modalities to effectively learning the key structure and schematic information of unseen instances. The feature aggregation module dynamically constructs local and global geometric information of the keypoints based on the keypoint foundation, thereby learning geometric information of unseen instances and improving the model's ability to learn geometric and structural information of unseen instances, as well as its generalization capability for pose estimation of unseen target instances. Our extensive experimental results on the REAL275 and CAMERA25 datasets demonstrate that the proposed keypoint enhancement method has excellent performance.

category level 6D pose estimation, key points enhancement, cross modal attention mechanism, adaptive weight fusion, adaptive dynamic neighborhood, sine position encoding, conditional normalization

doi: [10.1360/SST-2025-0356](https://doi.org/10.1360/SST-2025-0356)