基于双向特征融合的物体位姿估计方法

缪 君[1,2],严 杰[1],杜荣华[1],李 磊[3],储 珺[3]

(1. 南昌航空大学 航空制造与机械工程学院,南昌 330063;2. 中国科学院月球与深空探测重点实验室,北京 100190;3. 南昌航空大学 计算机视觉研究所,南昌 330063)

摘 要:为充分利用RGB图像的外观特征和深度图像的几何特征,提出了一种“外观-几何”特征并行融合的物体位姿估计方法。首先,在特征提取与融合阶段,构建了一种具有3个并行支流的双向融合体系结构,确保在每个编码层和解码层对并行的RGB图像特征和深度图像特征进行融合;同时,为避免重要特征丢失,且实现两种特征的充分融合,设计了两个互补的注意力机制,使两种特征获得局部和全局的互补;其次,在位姿推理计算阶段,考虑网络输出关键点与物体中心点之间的距离,提出了一种基于距离量和距离约束相结合的关键点检测网络,实现了精确的位姿估计。本文算法其在两个具有挑战性的6D物体位姿估计数据集上进行了测试,验证了其有效性。

关键词:位姿估计;双向特征融合;特征差异;距离约束;注意力机制

中图分类号:TP391.41 **文献标志码**:A **文章编号**:1671-5497(2026)02-0523-10

DOI:10.13229/j.cnki.jdxbgxb.20240748

A bidirectional feature fusion method for object position estimation

MIAO Jun[1,2],YAN Jie[1],DU Rong-hua[1],LI Lei[3],CHU Jun[3]

(1. *School of Aeronautical Manufacturing and Mechanical Engineering*, *Nanchang Hangkong University*, *Nanchang* 330063, *China*; 2. *Key Laboratory of Lunar and Deep Space Exploration*, *CAS*, *Beijing* 100190, *China*; 3. *Institute of Computer Vision*, *Nanchang Hangkong University*, *Nanchang* 330063, *China*)

Abstract: To fully leverage the appearance features of RGB images and the geometric features of depth images, this paper proposes an "appearance-geometry" features parallel fusion method for object position estimation. First, in the feature extraction and fusion stage, a three-stream bidirectional fusion architecture is constructed to ensure that the parallel RGB image features and depth image features are fused at each encoding layer and decoding layer. To prevent the loss of important features and achieve sufficient fusion of the two types of features, two complementary attention mechanisms are designed, enabling the two features to gain both local and global complementarities. Sercond, in the pose inference calculation stage,

收稿日期:2024-07-06.

基金项目:国家自然科学基金项目(62366032,62361043,62162045);中国科学院月球与深空探测重点实验室开放基金项目(LDSE202301).

作者简介:缪君(1979-),男,教授,博士.研究方向:三维场景理解与重建,工业视觉检测.
E-mail:miaojun@nchu.edu.cn

通信作者:杜荣华(1995-),男,讲师,博士.研究方向:深度学习,微小卫星在轨自主相对导航技术.
E-mail:7130l@nchu.edu.cn

considering the distance between the keypoints output by the network and the object's center point, a keypoint detection network based on a combination of distance metric and distance constraint is proposed, achieving accurate position estimation. The proposed algorithm has been tested on two challenging 6D object position estimation datasets, validating its effectiveness.

Key words: position estimation; bidirectional feature fusion; feature disparity; distance constraint; attention mechanism

0 引 言

物体位姿估计是计算机视觉领域中的关键技术,在机器人自动化操作、增强现实、自动驾驶等应用中发挥着重要作用,目前已成为国内外研究机构的热门研究课题[1,2]。然而,由于传感器噪声、光照变化、场景遮挡等现实因素,使得物体位姿估计问题仍面临巨大挑战。近年来,随着深度学习技术的迅速发展,利用卷积神经网络(Convolutional neural network, CNN)处理RGB图像解决这一问题成为研究热点[3]。然而,透视投影导致的几何特征丢失限制了该技术在相关场景中的应用。新型低成本RGB-D传感器的出现提供了额外的深度信息,可以解决这一问题,为特殊场景下的物体位姿估计问题提供了新颖的解决方案。

为克服光照条件差、场景遮挡等挑战,许多研究开始引入深度信息,以提供更精确的特征表示。Wang等[4]将深度图转换为点云特征,并与RGB图像中的颜色特征融合,预测目标最终的6D位姿。然而,该方法中RGB图像和点云特征的提取网络是独立设置的,无法实现信息共享和交流,导致特征融合不充分。He等[5]在网络每层添加点对像素和点对点融合模块,促进RGB-D信息的互补通信,但该方法在特征融合时可能导致重要信息丢失,尤其在遮挡情况下表现不佳。此外,特征提取与融合后,该方法通常依赖复杂的后处理操作实现最终的位姿估计,增加了计算成本和模型复杂性。相比之下,近期基于关键点检测的方法能直接、准确地定位物体位姿。Peng等[6]采用目标表面的关键特征点代替边界框的角点,以便更有效地获得上下文信息,但是二维图像上产生的误差转化到三维空间时易导致较大的定位偏差,这会影响目标最终的位姿估计精度。Lin等[7]提出了一种基于几何特征的端到端回归的位姿估计方法,将关键点偏移分解为单位向量和长度,并引入关键点采样策略,但关键点坐标的变化使网络无法充分学习其间的关系,导致泛化能力不足、鲁棒性较低。

针对上述物体位姿估计存在的各种问题,本文提出了一种"外观-几何"特征并行融合的物体位姿估计方法。为充分利用RGB图像和深度图像的特征信息,本文方法在网络每一层进行特征融合,并设计两个互补的注意力机制,以避免重要特征丢失,并实现特征的充分融合。此外,精确查找目标对象的关键点并建立对应关系对位姿估计至关重要。因此,本文提出了一种基于距离量和距离约束相结合的关键点检测网络,以提高物体关键点定位的准确性。

1 相关工作

1.1 基于RGB图像的物体位姿估计方法

基于RGB图像的物体位姿估计方法通常可以分为3类,即基于回归、基于模板和基于特征的方法。基于回归的方法[8]直接从RGB图像中输出位姿参数,但旋转空间的非线性限制了这类方法的推广。相反,基于模板的方法[9]需要在许多标记有真实位姿的模板中找到与目标最相似的模板,进而执行6D位姿估计。尽管这类方法对遮挡场景具有鲁棒性,但光照条件的剧烈变化会导致物体表面的颜色和亮度发生改变,易造成模板匹配失败。基于特征的方法[10]通过提取目标对象的2D关键点特征,以建立位姿估计所需的2D-3D对应关系。然而,透视投影会导致几何信息丢失,这种信息丢失限制了仅使用RGB图像处理方法的性能。

1.2 基于点云或深度图像的物体位姿估计方法

在工业物体位姿估计中,基于RGB图像的方法常因几何数据丢失、颜色和纹理信息不足而表现出局限性[11]。相比之下,基于点云和深度图像

的方法具有更大的优势。深度信息和点云信息包含丰富的形状几何特征,这对推断物体位姿具有重要意义。深度传感器和点云表示技术[12]的发展推动了基于点云的位姿估计方法的研究。这类方法利用3D ConvNet或点云网络进行特征提取和3D边界框预测,但点云信息的稀疏性和非纹理性降低了这类方法的性能。基于深度图像的方法[13]通过将深度图转换为点云数据,再利用点云数据中的几何信息和分割掩膜进行物体位姿估计,但分割掩膜的质量和准确性限制了这类方法的应用范围。此外,基于点云或深度图像的方法在处理物体表面反射光时,可能难以捕获实际点云数据,从而影响最终的物体位姿估计精度。因此,在物体位姿估计中,有必要充分考虑RGB图像的作用。

1.3　基于RGB-D图像的物体位姿估计方法

在解决物体位姿估计问题时,仅依赖RGB图像可能会受到光照条件差、低对比度、场景遮挡等挑战影响;而仅使用点云信息或深度信息,则面临数据结构化处理的困难。基于RGB-D数据的方法主要挑战在于如何充分利用RGB图像的外观特征和深度图像的几何特征。文献[14]采用两种特征提取方法分别处理RGB图像和深度图像,再将提取的特征进行融合,用于物体位姿估计,但该方法存在网络间信息无法共享的问题。因此,本文提出了一种“外观-几何”特征并行融合的物体位姿估计方法,使目标的外观特征和几何特征得到充分融合。这种设计不仅能保留RGB图像和深度分支的特征信息,还能充分利用融合分支的特征,最大限度地缩小两种特征之间的差异,有效解决了RGB图像的局限性。

2　算法框架

为充分融合RGB图像特征和深度图像特征的优势,本文提出了一种“外观-几何”特征并行融合的物体位姿估计算法,算法框架如图1所示,主要包括3个关键阶段:①RGB-D图像的语义分割阶段(图1左上部分)。该阶段旨在获取目标的边界框并进行裁剪,从而得到目标物体的颜色图和深度图,并将深度图转换为点云数据。②特征提取与融合阶段(图1右边部分)。该阶段构建了一种具有3个并行支流的双向融合体系结构,在每个编解码层均对RGB图像特征和深度图像特征进行融合,同时设计了两个互补的注意力机制,使目标的外观特征和几何特征获得局部和全局的互补。③位姿推理计算阶段(图1左下部分)。该阶段提出了一种结合距离量和距离约束的关键点检测网络,并通过最小二乘拟合方法求解物体的最

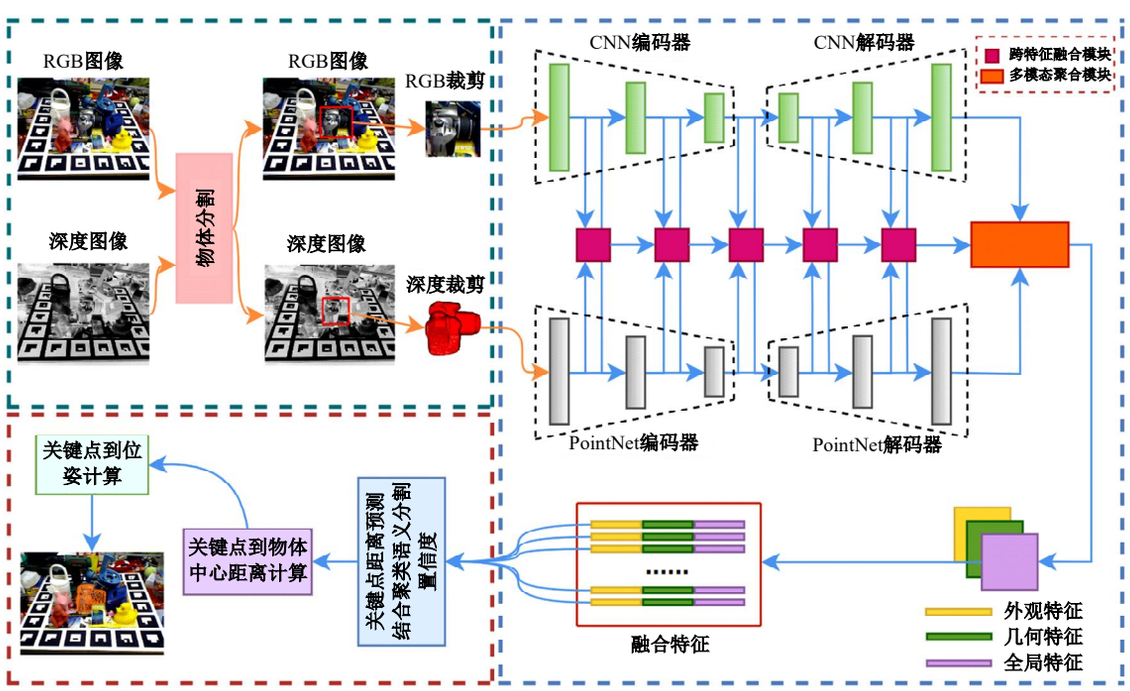

图1　本文提出的位姿估计网络结构图

Fig. 1　Structure diagram of the proposed pose estimation network

终位姿。

2.1 语义分割阶段

语义分割阶段的任务是从RGB-D图像中裁剪出目标对象,最终生成仅包含目标对象的彩色图像和深度图像。随着语义分割技术的成熟,本文采用语义分割网络[8]进行图像裁剪。该网络框架主要由编码器和解码器组成,输入图像经过编码器处理后提取特征信息,接着流入解码器,生成具有$N+1$个通道的语义分割图。其中,每个像素对应图像中的一个位置,并且每个通道的活跃像素表示图像中属于该类别的像素。

2.2 特征提取与融合阶段

由于RGB图像以颜色和纹理特征为主,可以突出物体的视觉轮廓边界;深度图像以空间三维特征为主,可以突出物体的几何边界。因此,不宜简单直接合并RGB图像特征和深度图像特征。为此,本文设计了一系列注意力机制模块。其中,通道注意力模块的结构如图2(a)所示。首先,输入特征图X通过全局平均池化提取空间信息,生成特征向量Z;其次,该向量经过卷积操作聚合通道间的关系,并采用Sigmoid函数归一化生成注意力权重;最后,这些权重与原始特征进行逐元素相乘,得到加权后的特征图。通过理解特征通道之间的相互关系,网络能够更准确地表达输入数据中的信息,提高模型的整体性能和准确性[15]。因此,本文利用这种方法设计了两种不同类型的特征处理模块,分别为跨特征融合模块和多模态聚合模块。

跨特征融合模块:在室内场景中,物体在颜色图和深度图中的特征信息表达不同,特征分布也各不相同。而通道注意力机制能使网络更多地关注目标对象特征丰富的区域,并过滤一些不重要的特征。为此,本文设计了跨特征融合模块(Cross feature fusion module, CFFM),如图2(b)所示,其中嵌入了两个通道注意力机制模块,分别处理RGB分支和深度分支的特征信息。

RGB图像处理采用ResNet18网络作为特征提取网络,输入图像通过卷积核大小为7×7、stride为2、padding为3的卷积层,以及归一化层和ReLU激活函数输出特征。将输出特征减少为原来的一半,并输入4个不同的残差模块,每个残差模块后都加入CCFM模块。其中,4层输出通道数分别为64、128、256和512,padding分别为0、2、1和1。

深度图像通过PointNet网络转换为点云数据,经过预处理后输入点云特征提取器和全局特征聚合器。点云特征提取器使用多层感知机(Multilayer perceptron, MLP)对每个点进行非线性变换,提取代表性特征。随后,将特征减少为原

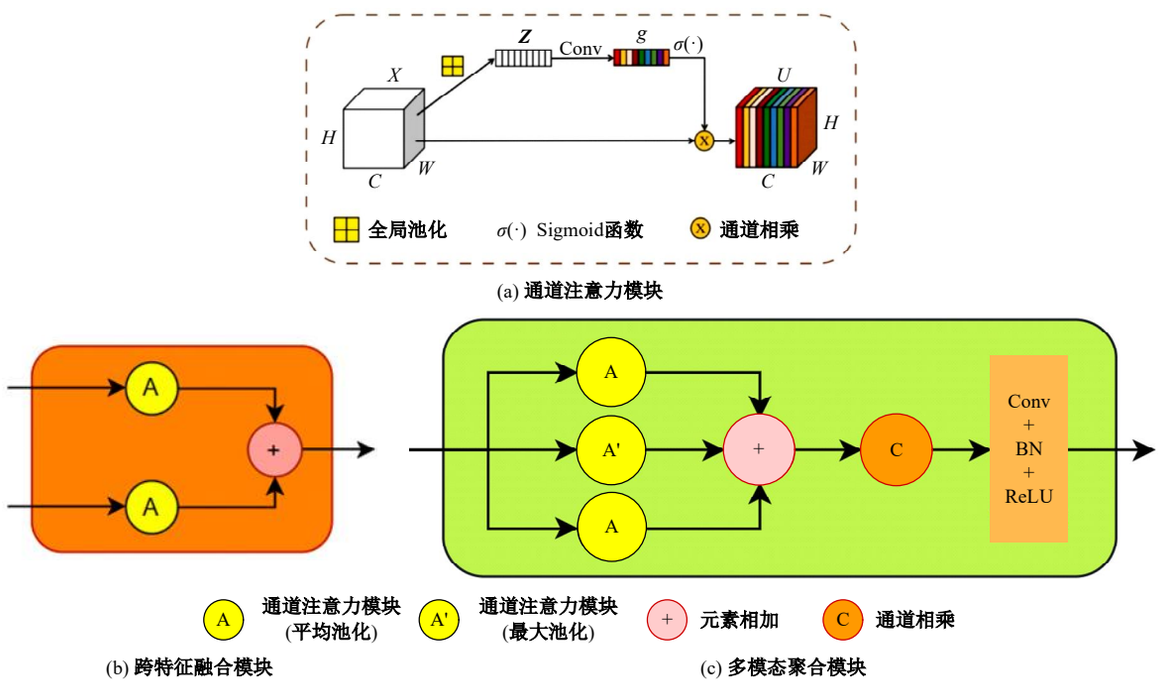

图2 本文提出的特征处理模块结构图

Fig. 2 The structure diagram of the proposed feature processing module

来的一半,输入CCFM模块,以增强特征的表达能力。最后,全局特征聚合器通过最大池化整合布局特征,确保全局特征包含所有点的关键信息。

在处理图像时,给定的RGB图像张量形状为$RGB_{in}\in\kappa^{C\times H\times W}$,深度图像张量形状为$Depth_{in}\in\kappa^{C\times H\times W}$,CFFM模块的操作可以描述为:

$$RGB_{per}=\rho\left(RGB_{in},Depth_{in}\right) \tag{1}$$

式中:ρ为CFFM融合模块;RGB_{per}为RGB分支和深度分支每一层的输出。

RGB分支和深度分支的每一层经过CFFM模块后输出为:

$$RGB_{out}=\left(RGB_{in}+Depth\right)/2 \tag{2}$$

$$Depth_{out}=\left(RGB_{in}+Depth_{in}\right)/2 \tag{3}$$

CCFM模块输出的结果将传播到编码器的下一层,以更准确、更高效地编码目标的外观特征和几何特征。每次融合结果减少到原来的一半,作为输入流入下一层编码器;另一半作为单独的融合分支,输出到下一层CCFM模块。这一过程降低了模型的复杂性,并使融合结果更加精细。

多模态聚合模块:在编码器将输出传送到解码器的过程中,模型无法捕捉所有重要的细节和信息。为了解决这一问题,本文设计了多模态聚合模块,如图2(c)所示。该模块基于通道注意力机制实现,结合全局平均池化和最大池化两个不同的通道注意力模块。其中,全局平均池化用于平滑整个通道特征,而最大池化则强调通道中最显著的特征。该模块接收来自RGB分支、深度分支和融合分支的输入特征,通过不同池化方法平衡对整体信息和局部显著特征的关注,从而增强网络对重要特征的捕捉能力,聚合更多特征信息。3个分支的输出特征通过元素相加融合,生成更全面的特征表示。随后,对融合后的特征进行逐元素相乘,进一步增强特征之间的关联性。最后,经过1×1卷积层、归一化和ReLU激活函数处理,输出最终的特征表示。

2.3　位姿推理计算阶段

在位姿推理计算阶段,通过处理特征提取与融合阶段的输出特征,获取最终的位姿信息。其主要内容包括:计算关键点到物体中心点之间的距离量,以及计算关键点到最终的物体位姿信息,如图3所示。本文构建了关键点与物体中心距离量预测模块、聚类语义分割模块、置信度模块3个不同的任务模块,对关键点进行处理,并利用约束条件计算关键点到物体中心点之间的距离。在相机坐标系下,从检测到的关键点中选取最佳的3个点,并与物体坐标系对应关键点坐标进行匹配。最后,通过最小二乘拟合方法计算,获取最终的物体位姿信息。

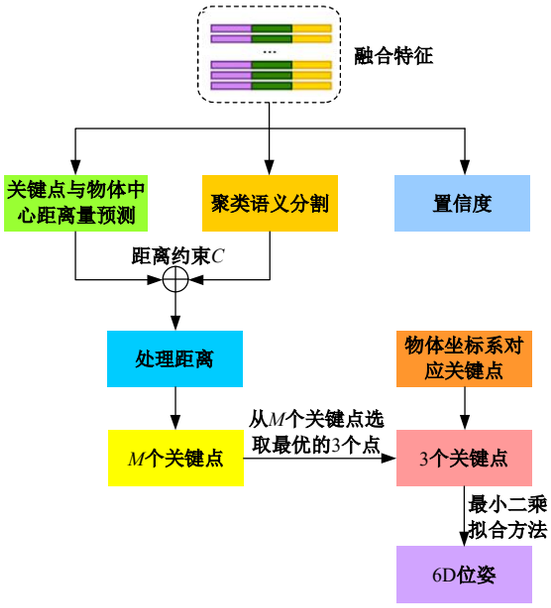

图3　本文提出的关键点检测网络(DCKP)结构图

Fig. 3　Structure diagram of the proposed keypoint detection network (DCKP)

2.3.1　计算关键点到物体中心点之间距离量

(1)关键点与物体中心距离量预测模块

在经过特征提取和融合阶段后,该模块利用MLP预测关键点与物体中心的距离。提取到的特征信息经过网络处理后,产生一个具有M个通道的输出,每个通道包含表示关键点到物体中心点的距离值。关键点与物体中心点距离量的监督采用$L_{keypoint}$损失,具体表达式如式(4)所示:

$$L_{keypoint}=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\left\| d_i^j-d^{*j}_i\right\| \tag{4}$$

式中:M为所选目标关键点的总数;N为物体中心点的总数;d_i^j为第i个关键点与第j个物体中心点的预测距离;d^{*j}_i为第i个关键点与第j个物体中心点的实际距离。

(2)聚类语义分割模块

当场景中存在多个物体时,传统方法通常通过目标检测或者分割技术对图像进行预处理,以获得物体的边界框。而通过中心点信息确定物体

的大致位置和边界,并进一步优化语义分割边界,可以提高分割效果的准确性和鲁棒性。本文采用MLP完成两项预测任务:一是预测N个中心点的语义分割,二是预测相对于物体实际中心点的偏移量Δs。然而,在对同一物体不同点进行预测时,训练过程中易出现噪声等问题,导致预测结果与真实值之间存在偏差。为了解决这一问题,使用聚类算法集中处理同一类别的预测中心点,从而确定每个物体的最终中心点位置,消除噪声引起的错误点影响。在这一任务中,用于语义分割的损失函数定义为:

$$L_{\text{semantic}}=-\alpha\left(1-\beta_i\right)^{\gamma}\log\left(\beta_i\right) \tag{5}$$

式中:α为平衡样本在损失函数中的贡献;γ为调节损失函数对样本的关注度;β_i为模型对该点预测的正确性。

物体中心点预测的损失函数定义为:

$$L_{\text{center}}=\frac{1}{N}\sum_{i=1}^{N}\left\|\Delta s_i-\Delta s_i^*\right\| \tag{6}$$

式中:Δs_i为预测的中心点偏移量;Δs_i^*为中心点偏移量的实际值。

考虑到关键点的位置会随着物体的旋转、平移等位姿变化而改变,为确保预测的关键点到物体中心的距离尽可能接近实际距离,本文提出了一种约束条件C,以确保关键点与物体中心点距离量预测的损失函数L_{keypoint}和聚类语义分割的中心点预测损失函数L_{center}的一致性。距离约束条件定义为:

$$C=\lambda\sum_{i=1}^{N}\sum_{j=1}^{M}\max\left(0,\left\|\Delta s_i-\Delta s_i^*-\left(d_i^j-d^{*j}_i\right)\right\|-\varepsilon\right) \tag{7}$$

式中:λ为控制约束条件对损失函数影响的超参数;$\max\left(0,\left\|\Delta s_i-\Delta s_i^*-\left(d_i^j-d^{*j}_i\right)\right\|-\varepsilon\right)$为确保预测偏移与实际偏移之间的差异在允许范围内;$\varepsilon$为容差参数。

(3)置信度模块

当物体外观与背景相似时,关键点的预测易出现错误,从而导致位姿计算引入误差,影响最终的位姿估计性能。为解决这一问题,该模块利用MLP预测关键点的位置,并输出关键点的置信度,以评估预测结果的准确性。通过预测关键点位置与真实位置之间的误差,设定置信度的阈值为0.8,可以有效识别和排除错误的预测点。其中,用于预测关键点的损失函数定义为:

$$L_{\text{conf}}=\frac{1}{M}\sum_{i=1}^{M}\left\|y_i\cdot\lg c_i+\left(1-y_i\right)\lg\left(1-c_i\right)\right\| \tag{8}$$

式中:y_i为置信度标签;c_i为预测的置信度。

由式(8)可以看出,置信度阈值越大,损失越小,同时这也反映了关键点的预测值与实际值之间的接近程度。

2.3.2 计算关键点到最终的物体位姿信息

根据2.3.1节,通过计算关键点到物体中心之间的距离,求解相机坐标系下的关键点坐标P_c,并与物体坐标系下对应的点坐标P_i进行拟合。通过最小二乘拟合方法处理已知的三维点对,构建并最小化损失函数$L_{\min}$,可以计算物体位姿的旋转矩阵$\boldsymbol{R}$和平移矩阵T。然而,实际情况中易存在严重遮挡,导致某些关键点无法观测。因此,通过筛选高精度关键点,可提升物体位姿估计的准确性。

由于最小二乘拟合方法至少需要3个关键点进行位姿计算,为选择最合适的3个关键点,可以通过暴力搜索法从M个关键点中选取任意3个点进行组合解算,总共有C_M^3种组合。每种组合均可求解得到1个候选位姿,通过选择与平移变换不同的候选位姿,最终选择具有最小误差的候选位姿作为物体的最终位姿估计。该方法在关键点数量较少的情况下,通过有效的组合选择和误差计算,实现了对物体位姿的高效求解。其中,通过最小化平方损失求解$\boldsymbol{R}$和T,该方法的具体表达式为:

$$L_{\min}=\sum_{i=1}^{M}\left\|P_i-\left(\boldsymbol{R}\cdot P_c+T\right)\right\|^2 \tag{9}$$

3 实验分析

3.1 数据集

(1)LineMOD数据集:由Hinterstoisser等[16]创建,该数据集是由13个无纹理3D对象实例和13个注册视频序列组成的大型数据集,具有场景杂乱、光照变化大、物体纹理弱等挑战。

(2)YCB-Video数据集:由Calli等[17]创建,该数据集由92个RGB-D视频组成,包含21个不同的对象实例,具有光照变化大、图像噪声不均匀、场景遮挡等挑战。在训练时,使用80 000张合成图像进行模型训练和评估。

3.2　评估指标

物体位姿估计通常采用平均距离度量ADD和ADDS两项指标进行评估。对于非对称物体，采用ADD指标。该指标计算由预测位姿和地面真实位姿变换的对象顶点之间的点对平均距离表示，具体表达式如式(10)所示。对于对称物体，采用ADDS指标，该指标计算由预测位姿和地面真实位姿变换的对象顶点之间的最近点对平均距离表示，具体表达式如式(11)所示。

$$D_{\mathrm{ADD}}=\frac{1}{m}\sum_{v\in o}\left\|(\boldsymbol{R}x+t)-(\boldsymbol{R}^{*}x+t^{*})\right\| \quad (10)$$

$$D_{\mathrm{ADDS}}=\frac{1}{m}\sum_{v\in o}\left\|(\boldsymbol{R}x_{1}+t)-(\boldsymbol{R}^{*}x_{2}+t^{*})\right\| \quad (11)$$

式中：x为对象网格o上的顶点；o为对象网格；m为对象网格o的顶点总数；$\boldsymbol{R}$、T为模型预测的位姿；$\boldsymbol{R}^{*}$、T^{*}为地面真实位姿。

3.3　实验环境

本文算法的运行环境为Ubuntu18.04操作系统、Pytorch1.6.0平台、Python3.8语言。模型的初始学习率为10^{-4}，权重衰减率为10^{-5}，Batch Size为8。训练阶段采用1块运行内存为24G的NVDIA RTX 3090TI GPU显卡，共训练400个轮次。

3.4　消融实验分析

为了验证本文提出的“外观-几何”特征并行融合物体位姿估计方法对整个网络的影响，本文在LineMOD数据集和YCB-Video数据集上进行了消融实验。

3.4.1　LineMOD数据集的实验结果分析

表1展示了本文方法在LineMOD数据集上，基于ADD(S)评估指标度量位姿估计的结果。由表1可知，本文方法在LineMOD数据集上的准确率达到了99.4%。同时，将本文方法与另外3种同样使用RGB-D图像作为输入的方法进行比较。

本文方法(KP)采用与PVN3D和FFB6D方法相同的关键点检测网络进行训练，评估结果的准确性达到98.6%。其中，较同样采用关键点检测网络的PVN3D和FFB6D方法分别提升7.3、2.8百分点，较采用后处理(Iterative closest point，ICP)的DenseFusion方法提升13.3百分点。这表明本文提出的“外观-几何”特征并行融合方法能够有效地融合RGB图像和深度图像的特征。LineMOD数据集的可视化结果如图4所示，可以看出，本文方法相比于其他模型方法在结构上位姿估计更加精确。即使在像电钻(driller)这样存在严重遮挡的场景中，本文方法仍能取得优异的性能。

为了验证本文提出的基于距离量和距离约束相结合的关键点网络对整个模型的影响，本文在广泛应用于物体识别和位姿估计的LineMOD数据集上进行了消融实验，以评估所提关键点检测网络的有效性及其对整体模型的具体影响。本次消融实验仍然采用上述两项评价指标进行定量分

表1　在LineMOD数据集上，本文方法与其他方法对比

Table 1　Comparison of the proposed method with other methods on the LineMOD dataset　%

	RGB-D									
	DenseFusion		PVN3D		FFB6D		本文方法(KP)		本文方法	
object	ADD	ADDS	ADD	ADDS	ADD	ADDS	ADD	ADDS	ADD	ADDS
ape	83.9	70.2	90.9	84.6	96.0	80.5	98.4	90.6	**99.8**	**96.5**
benchvise	86.9	83.1	93.1	89.3	96.1	94.8	98.3	94.6	**98.5**	**95.1**
camera	84.2	88.4	94.3	94.2	97.4	96.3	98.6	96.6	**99.6**	**98.6**
can	86.0	76.2	90.5	84.8	**98.2**	88.5	97.6	89.6	97.9	**96.8**
cat	90.4	81.0	95.6	87.5	97.5	96.2	99.8	97.0	**99.8**	**97.6**
driller	88.0	70.7	91.7	83.8	96.0	89.3	98.8	93.9	**99.6**	**97.0**
duck	84.1	82.7	94.3	87.1	97.1	95.7	99.1	94.6	**99.3**	**97.8**
eggbox	87.2	85.2	92.9	89.5	97.7	96.1	98.1	96.9	**99.4**	**99.4**
gule	83.5	74.5	88.2	81.6	93.3	88.6	98.7	94.1	**99.7**	**99.7**
holepuncher	86.0	82.3	94.8	81.0	96.6	93.7	99.2	**95.5**	**99.5**	95.3
iron	87.0	83.6	89.5	87.3	97.4	96.5	**99.6**	96.9	99.2	**97.3**
lamp	81.6	75.3	88.4	79.8	96.0	93.2	98.8	96.3	**99.5**	**97.7**
phone	79.6	79.6	82.3	88.3	86.0	90.2	96.3	96.3	**99.6**	**98.6**
ALL	85.3	79.4	91.3	86.1	95.8	92.3	98.6	94.8	**99.4**	**97.5**

注：表中粗体值为最佳值。

图4 在LineMOD数据集上,本文方法与其他方法可视化结果

Fig. 4 Visualization results of the proposed method with other methods on the LineMOD dataset

析,结果如表2所示。

由表2可知,本文提出的基于距离量和距离约束相结合的关键点检测网络在物体位姿估计中表现更优。在特征提取和融合网络提取到一致性特征信息的基础上,分别采用后处理方法和关键点检测方法估计物体的位姿信息。与这两种物体位姿估计方法相比,本文算法的物体位姿估计精度更优。在位姿推理计算过程中,本文方法能考虑网络直接输出的关键点与物体中心点之间的距离,以确保预测关键点与物体中心点保持正确的空间关系,从而实现精确的物体位姿估计。

表2 本文所提关键点检测网络的有效性验证

Table 2 Validation of the effectiveness of the keypoint detection network proposed in this paper %

方法	Baseline+ICP	Baseline+KP	Baseline+DCKP
ADD	93.0	95.5	**97.4**
ADDS	86.9	91.8	**94.5**

注:表中粗体值为最佳值。

3.4.2 YCB-Video数据集的实验结果分析

表3展示了本文方法在YCB-Video数据集上,基于ADD(S)评估指标度量物体位姿估计的结果。本文方法在YCB-Video数据集上的物体位姿估计精度达到了98.8%,并与两种不同输入数据类型的方法进行了比较。其中,带*标记的类别对象为对称物体,采用ADDS指标评估;其他物体则采用标准的ADD指标评估。本文算法的评估结果显著优于其他4种算法,在大多数物体对象上取得最优值,表明其在光照变化、场景遮挡等场景中具有优异的适用性和性能。此外,图5

表3 在YCB-Video数据集上,本文方法与其他方法对比

Table 3 Comparison of the proposed method with other methods on the YCB-Video dataset %

	RGB				RGB-D					
	PoseCNN		PVNet		DenseFusion		FFB6D		本文方法	
object	ADD	ADDS	ADD	ADDS	ADD	ADDS	ADD	ADDS	ADD	ADDS
002_master_chef_can	83.9	50.2	90.9	74.6	95.3	70.7	96.3	80.6	**98.2**	**93.2**
003_cracker_box	76.9	53.1	87.1	79.3	92.5	86.9	96.3	94.6	**99.8**	**96.3**
004_sugar_box	84.2	68.4	94.3	84.2	95.1	90.8	97.6	96.6	**99.2**	**97.9**
005_tomato_soup_can	81.0	66.2	90.5	79.8	93.8	84.7	95.6	89.6	**97.8**	**94.8**
006_mustard_bottle	90.4	81.0	90.6	83.5	95.8	90.9	97.8	97.0	**98.8**	**98.5**
007_tuna_fish_can	88.0	70.7	91.7	73.8	95.7	79.6	96.8	**89.7**	**99.8**	88.9
008_pudding_box	79.1	62.7	89.3	84.1	94.3	89.3	97.1	94.6	**97.7**	**96.8**
009_gelatin_box	87.2	75.2	92.9	89.5	97.2	95.8	**98.3**	96.9	98.1	**97.3**
024_bowl*	69.6	69.6	80.3	80.3	86.0	86.0	96.3	96.3	**99.8**	**99.8**
025_mug	78.2	58.5	90.7	76.6	95.3	83.8	97.3	**94.2**	**98.5**	93.4
035_power_drill	72.7	55.3	87.4	78.4	92.1	83.7	97.2	95.9	**98.0**	**97.6**
036_wood_block*	64.3	64.3	84.2	84.2	89.5	89.5	92.6	92.6	**98.9**	**98.9**
037_scissors	56.9	35.8	84.2	70.3	90.1	77.4	97.7	**95.7**	**98.5**	95.2
040_large_maker*	71.7	58.3	89.5	81.0	95.1	89.1	96.6	89.1	**99.0**	**99.0**
051_large_clamp*	50.2	50.2	63.6	63.6	71.5	71.5	96.8	96.8	**99.1**	**99.1**
061_foam_brick	88.0	88.0	83.1	83.1	92.2	92.2	97.3	97.3	**98.6**	**98.6**
ALL	76.4	63.0	86.9	79.1	92.0	85.1	96.7	93.6	**98.8**	**96.6**

注:表中粗体值为最佳值,带*标记的类别对象为对称物体。

图 5　在 YCB-Video 数据集上,本文方法训练的可视化结果

Fig. 5　Visualization results of the proposed method trained on the YCB-Video dataset

展示了 YCB-Video 数据集的可视化结果,可见本文方法成功实现了对场景中不同类别物体准确的位姿估计。

4　结束语

本文提出了一种“外观-几何”特征并行融合的物体位姿估计方法,主要用于估计 RGB-D 图像中已知物体的 6D 位姿。在特征提取和融合阶段,构建了一种具有 3 个并行支流的双向融合结构,并设计了两个互补的注意力机制,有效解决了特征融合不充分的问题。此外,在位姿计算阶段,提出了一种基于距离量和距离约束相结合的关键点检测网络,从而提高了物体位姿推理估计的精度。实验结果证明:本文方法可以在严重遮挡、低对比度和光照条件差的场景中准确估计目标物体的位姿信息。在两个具有挑战性的 6D 物体位姿估计基准数据集上的评估结果显示,本文方法明显优于现有方法,验证了其在物体位姿估计任务中的先进性和有效性。然而,尽管本文方法在许多场景下表现优异,但在处理极其复杂的背景时,其鲁棒性仍需进一步提高。此外,本文方法对训练数据的多样性存在依赖,易导致在未知物体上的性能下降。未来研究可集中于增强模型的鲁棒性,探索无监督和半监督学习策略,以减少对标注数据的需求,并优化模型结构以提高实时性。

参考文献:

[1] Guan J, Hao Y M, Wu Q X, et al. A survey of 6DoF object pose estimation methods for different application scenarios[J]. Sensors, 2024, 24(4): 1076.

[2] Marullo G, Tanzi L, Piazzolla P, et al. 6D object position estimation from 2D images: A literature review [J]. Multimedia Tools and Applications, 2023, 82 (16): 24605-24643.

[3] 王静, 金玉楚, 郭苹, 等. 基于深度学习的相机位姿估计方法综述[J]. 计算机工程与应用, 2023, 59(7): 1-14.
Wang Jing, Jin Yu-chu, Guo Ping, et al. A review of camera pose estimation methods based on deep learning[J]. Computer Engineering and Applications, 2023, 59(7): 1-14.

[4] Wang C, Xu D E, Zhu Y K, et al. Dense Fusion: 6D object pose estimation by iterative dense fusion [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 3343-3352.

[5] He Y S, Huang H B, Fan H Q, et al. FB6D: A full flow bidirectional fusion network for 6D pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEEE, 2021: 3003-3013.

[6] Peng S D, Liu Y, Huang Q X, et al. PVNet: Pixel-wise voting network for 6DoF pose estimation[J]. Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 3212-3223.

[7] Lin S F, Wang Z R, Ling Y G, et al. E2EK: End-to-end regression network based on keypoint for 6D pose estimation[J]. IEEE Robotics and Automation Letters, 2022, 7(3): 6526-6533.

[8] Xiang Y, Schmidt T, Narayanan V, et al. Pose CNN: A convolutional neural network for 6D object pose estimation in cluttered scenes[J]. ArXiv Preprint, 2017, 11: 171100199.

[9] Zakharov S, Shugurov I, Ilic S. DPOD: 6D pose object detector and refiner[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2019: 1941-1950.

[10] 王连明, 吴鑫. 基于姿态估计的物体 3D 运动参数测量方法[J]. 吉林大学学报:工学版, 2023, 53(7): 2099-2108.
Wang Lian-ming, Wu Xin. Measurement of 3D motion parameters of an object based on attitude estimation[J]. Journal of Jilin University (Engineering and Technology Edition), 2023, 53(7): 2099-2108.

[11] Ding Z F, Sun Y X, Xu S J, et al. Recent advances and perspectives in deep learning techniques for 3D point cloud data processing[J]. Robotics, 2023, 12(4): 100.

[12] Zhou J, Chen K, Xu L L, et al. Deep fusion transformer network with weighted vector-wise keypoints voting for robust 6D object pose estimation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2023: 13967-13977.

[13] 白琳, 刘林军, 李轩昂, 等. 基于自监督学习的单目图像深度估计算法[J]. 吉林大学学报:工学版, 2023, 53(4): 1139-1145.
Bai Lin, Liu Lin-jun, Li Xuan-ang, et al. A depth estimation algorithm for monocular images based on self-supervised learning[J]. Journal of Jilin University (Engineering and Technology Edition), 2023, 53(4): 1139-1145.

[14] Song C, Song J R, Huang Q X. HybridPose: 6D object pose estimation under hybrid representations[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 431-440.

[15] 张宸嘉, 朱磊, 俞璐. 卷积神经网络中的注意力机制综述[J]. 计算机工程与应用学报, 2021, 57(20): 64-72.
Zhang Chen-jia, Zhu Lei, Yu Lu. A review of attention mechanisms in convolutional neural networks[J]. Journal of Computer Engineering & Applications, 2021, 57(20):64-72.

[16] Hinterstoisser S, Lepetit V, Ilic S, et al. Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes[C]// Computer Vision-ACCV 2012: 11th Asian Conference on Computer Vision. Piscataway, NJ: IEEE, 2013: 548-562.

[17] Calli B, Singh A, Walsman A, et al. The YCB object and model set: Towards common benchmarks for manipulation research[C]// International Conference on Advanced Robotics. Piscataway, NJ: IEEE, 2015: 510-517.