ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 5,Sep. 2024,pp. 1214－1227
DOI:10. 16337/j. 1004-9037. 2024. 05. 013

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于双向融合纹理和深度信息的目标位姿检测

张亚炜，付东翔

(上海理工大学光电信息与计算机工程学院,上海 200093)

摘　要：针对在硬件设备资源有限的情况下,深度相机在非结构化场景如何获取物体精确的位姿信息问题,提出一种基于双向融合纹理和深度信息的目标位姿检测方法。在学习阶段,两个网络采用全流双向融合(FFB6D)模块,纹理信息提取部分引入轻量的 Ghost 模块,减少了网络的计算量,并加入能增强有用特征的注意力机制CBAM,深度信息提取部分扩展了局部特征并多层次特征融合,获取更全面的特征;在输出阶段,为提高效率利用实例语义分割结果过滤背景点,再进行3D关键点检测,最终通过最小二乘拟合算法得到位姿信息。在LINEMOD、Occlusion LINEMOD和YCB-Video公共数据集上验证,其精度分别达到了99.8%、66.3%和94%,且参数量减少了31%,表明改进的位姿估计方法在保证精度的同时,也减少了参数量。

关键词：双向融合;Ghost;注意力机制;深度学习;位姿估计
中图分类号：TP391.41　　**文献标志码**:A

Target Position Detection Based on Bidirectional Fusion of Texture and Depth Information

ZHANG Yawei, FU Dongxiang

(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Aiming at the problem of how to obtain accurate positional information of objects in unstructured scenes by depth cameras with limited hardware device resources, a target position detection method based on bidirectional fusion of texture and depth information is proposed. In the learning phase, two networks adopt the full-flow bidirectional fusion (FFB6D) module, the texture information extraction part introduces the lightweight Ghost module to reduce the computation of the network, and adds the attention mechanism CBAM that can enhance useful features, and the depth information extraction part extends the local features and multilevel feature fusion to obtain more comprehensive features. In the output stage, in order to improve the efficiency, the instance semantic segmentation results are utilized to filter background points, then 3D keypoint detection is performed, and finally the position information is obtained by the least square fitting algorithm. Validations are carried out on LINEMOD, Occlusion LINEMOD and YCB-Video public datasets, whose accuracies reach 99.8%, 66.3% and 94%, respectively, and the amount of parameters is reduced by 31%, showing that the improved position estimation method can canreduce the number of parameters while guaranteeing the accuracy.

基金项目:国家自然科学基金(61703277)。
收稿日期:2023-08-22;**修订日期**:2023-12-21

Key words: bidirectional fusion; Ghost; attention mechanism; deep learning; position estimation

引 言

近年来,机器人技术在各领域中有着重要作用,如工业、医疗和服务行业等领域的应用,视觉的加入让这技术更好地运用到实际应用中[1]。通过传感器获取视觉信息,再将视觉信息传入控制系统,引导机器人对环境进行理解,提高机器人的智能化。随着机器视觉技术的不断发展,基于视觉引导的物体拾取操作、自动驾驶及精确定位导航等[2]诸多实际应用中都需要对目标物体进行6D位姿估计。6D位姿估计[3]是计算物体的世界坐标系(或参考坐标系)与相机坐标系之间的转换关系,即两个坐标系之间实现坐标转换的三维旋转矩阵和三维平移向量,根据该转换关系,将视觉中目标位置等信息转换为世界坐标系(参考坐标系)的位置信息,以便于对物体操作的后续处理。若缺少对目标的位姿估计,机器人将无法对目标进行精确的操作,因此位姿估计是机器视觉技术应用重要的组成部分。

位姿估计方法可分为传统方法和深度学习方法。传统方法主要包括基于模板匹配和基于对应的方法[4]。基于模板匹配法是从物体的二维图像或深度图中提取纹理特征,获取不同角度的模板,形成标注信息的模板库,检测到的目标物体与之匹配到最相似的模板作为物体的位姿,如Hinterstoisser等[5]提出的LINEMOD算法;基于对应的方法是将图像和三维模型之间的特征进行匹配,建立图像中目标物体与目标模型的对应关系,根据对应关系来获得6D位姿,常用的特征提取算法有SIFT[6]、FPFH[7]、SHOT[8]等,获取检测图像与模型之间的对应关系可使用PnP算法恢复位姿或全局配准优化位姿。然而,这些传统的方法太过依赖物体的几何形状和纹理信息,若在复杂的场景下有较大的影响,难以准确估计出目标物体的位姿。

随着机器学习和深度学习的不断发展,基于深度学习的位姿估计方法变得流行起来。Kehl等[9]扩展SSD直接回归6D位姿,Tekin等[10]提出的YOLO6D直接检测到3D边界框顶点的2D投影并通过PnP算法估计位姿。由于旋转空间的非线性,这些方法泛化能力不强,于是Peng等[11]提出了PVNet,通过霍夫投票得到关键点再用PnP计算位姿,但由于投影只能对应3D点信息,而这些点之间的连接关系无法获取使得几何信息丢失,从而限制了性能。随着3D相机的出现,能够轻易获取RGB-D数据,从深度相机中获取的额外深度信息,为位姿估计提供了更多的有用信息,Wang等[12]引入DenseFusion框架,设计了一种稠密的像素级融合方式,将RGB特征和点云特征以一种更合适的方式进行了整合。随后许多2D算法也被扩展到3D,并获得了良好的性能,PVN3D[13]是第一个将2D关键点方法扩展到3D关键点定位的方法,该方法使用3D关键点霍夫投票,然后使用最小二乘拟合算法来预测6D姿态。FFB6D[14]是在PVN3D基础上的改进方法,在学习阶段引入一个双向融合模块,将融合应用于每个编码和解码层,以获得更准确的效果。ICG框架[15]提出了一种结合区域和深度信息的概率跟踪器,仅依赖于物体几何形状,然而点云数据本身具有稀疏性,缺乏足够的纹理信息,这对这些方法的性能造成了限制。之后又提出ICG+算法[16]加入了额外的纹理模式,用于灵活的多相机信息融合。BiCo-Net[17]提出了一种新的双向对应映射网络,先在典型位姿回归的指导下生成点云,从而结合位姿敏感信息来优化局部坐标及其法向量的生成。然而现有方法在提取RGB-D数据特征信息方面仍然面临挑战,其并行结构带来大量的参数,且图像中采样大量的点,数据处理量较大,算法效率有待提高。

针对上述位姿估计方法中参数量大、算法效率不足等问题,本文提出一种双向融合纹理和深度信息的目标位姿检测方法。具体做法如下:

(1)在纹理特征提取部分引入轻量的Ghost模块来减轻模型,更好地在嵌入式设备上进行部署。

(2)为保证其精度,在骨干网络中分别对RGB部分和点云部分加入注意力机制和多层次特征融合

模块，充分提取纹理信息和深度信息。

（3）通过滤波背景点算法对背景点设置权重，利用实例语义分割结果对其进行过滤，降低点的数量，使得位姿检测响应速度更快，提高速度。

1 目标位姿检测模型

1.1 网络总体结构

本文双向融合纹理和深度信息的目标位姿检测方法是基于全流双向融合FFB6D改进的，网络结构如图1所示。给定RGB-D图像作为输入，同时使用已有的相机内参将深度图像转换为点云。全卷积网络部分用于提取纹理信息，引入Ghost模块，替换原先预训练的ResNet34[18]中的普通卷积，并采用能采集空间与通道注意力信息的卷积注意力模块（Convolutional block attention module，CBAM）[19]，通过关注局部区域实现增强特征，提升识别精度；点云网络从点云中提取深度信息，采用的是轻量级高效的点云网络RandLA-Net[20]，该网络用随机采样算法来减少点云采样的计算量，但会丢失一些关键点云，提出改进的局部特征聚合模块DLFA，通过扩张K最近邻（K nearest neighbor，KNN）搜索[21]在降低学习复杂度的同时还提升网络的感受野，并进行多层次特征融合（Multi-level hierarchical feature fusion，MHFF），将同一级别的特征从下到上连接起来补充深度信息。在表示学习阶段，在两个网络的全流程中构建了双向融合模块（FFB fusion），将融合模块应用于每个编码、解码层，使两个网络的局部和全局互补信息获得更好的表示，将提取的逐点特征被馈送到3D关键点检测模块和实例语义分割模块，用实例语义分割的结果通过滤波背景点算法过滤背景点来优化位姿估计，获得单个图像中的每个对象的3D关键点后，通过聚类算法对关键点投票，再采用最小二乘拟合算法预测关键点，最后给出6D位姿估计参数。

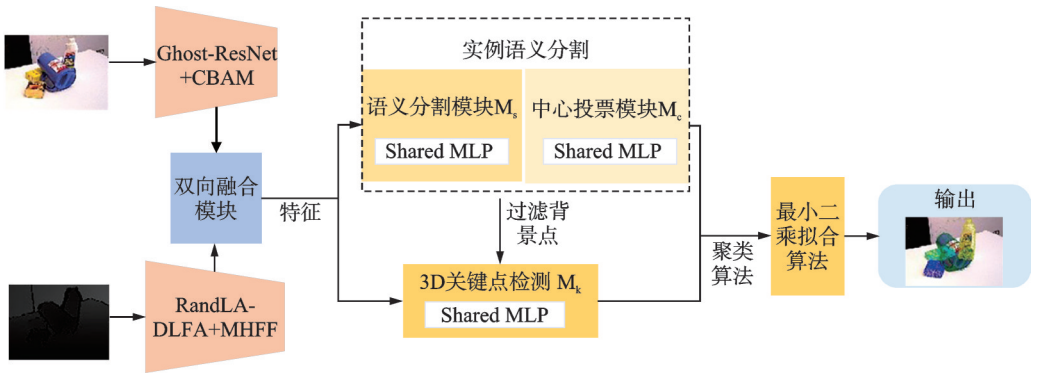


图1 目标位姿检测网络结构图

Fig.1 Structure of target position detection network

1.2 特征提取网络

本文给定的输入为RGB-D图像，需要用两种网络分别对纹理和深度信息进行特征提取，并将获取的纹理信息与深度信息充分融合。首先，对RGB-D的两种信息分别进行处理，在处理深度图时利用相机内参将其转化为点云进行，然后送入ResNet34-PSPNet[22]和RandLA-Net的特征提取网络分别对RGB图和点云进行特征提取。在特征融合部分，若仅在最后进行简单的特征拼接来实现信息融合，将会缺失部分特征；与原来的方法一样，在每个编码层和解码层都进行了纹理和几何信息的融合，实现特征互补，如图2所示，上部分是CNN网络，下部分是PCN网络，用双向融合方式将两者融合起来。其中PSPModule指PSPNet的金字塔池模块，DLFA为扩展局部特征聚合模块。

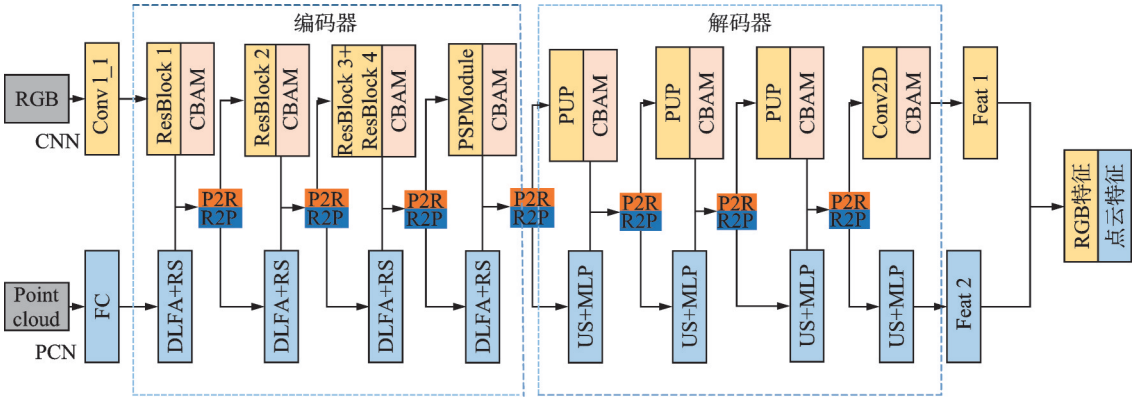


图2 特征提取网络结构图

Fig.2 Structure of feature extraction network

每个融合阶段都进行像素到点融合(R2P)模块和点到像素融合(P2R)模块,由于给定的RGB-D图像是对齐良好的,可以将3D点云作为连接像素和点特征的桥梁,用相机固有矩阵将每个像素的深度提升到其相应的3D点,并获得与RGB图对齐的XYZ图。像素到点融合模块是将RGB特征融合到点云特征,如图3(a)所示,将每个点特征在XYZ图找到它的最近邻$\boldsymbol{K}_{\mathrm{R2P}}$,并找到与之对应的像素特征,通过Max Pooling和Shared MLP来压缩到与点云特征相同大小,再通过Shared MLP融合外观特征和几何特征,最终获得融合的RGB特征。点到像素融合模块是将每个像素在XYZ图上的特征,在点云特征里找到其对应的最近邻$\boldsymbol{K}_{\mathrm{P2R}}$,对点云特征进行压缩和最大池化,最后点云特征连接相应的RGB特征,再用Shared MLP生成融合的点云特征,如图3(b)所示。

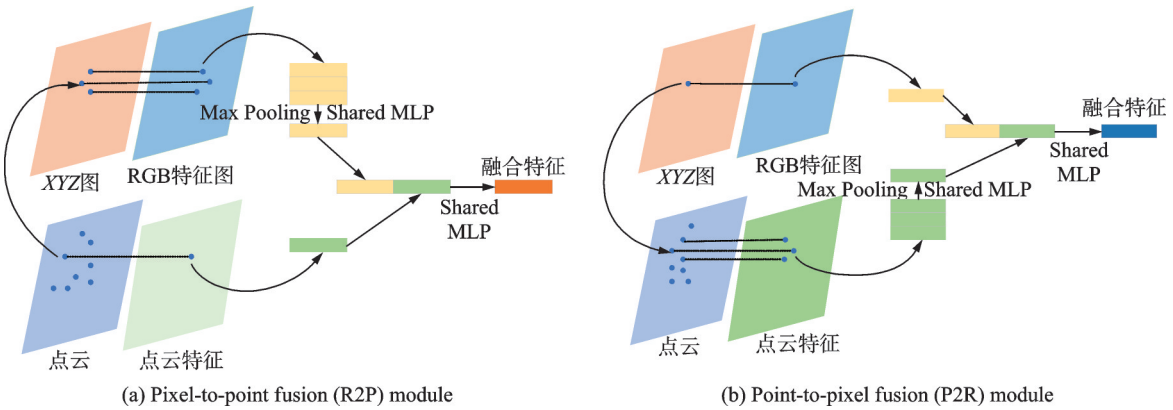


图3 RGB和点云特征融合结构图

Fig.3 Structure of RGB and point cloud feature fusion

1.2.1 纹理特征提取网络

纹理特征提取部分采用的是ResNet34-PSPNet这种编码器-解码器结构,在RGB图像编码器部分由5个卷积层和PSPNet中金字塔池化模块组成,解码器部分由3个上采样模块和PSPNet最后的卷积层组成。由于设备资源有限,深度网络模型往往模型过大且计算量太大,很难在嵌入式设备上部署,这里从减少冗余特征图来看,本文引入轻量化网络GhostNet[23]中的Ghost模块,来增强了网络的学习能力并减小了网络模型,减少运行时间。

Ghost模块是一个即插即用的新模块，可以插在纹理特征提取网络中搭建出轻量级的网络Ghost-ResNet。Ghost模块核心思想是设计一种分阶段的卷积计算模块，在少量的非线性卷积得到的特征图基础上再进行一次线性卷积获得Ghost特征图，以此消除冗余特征来获得更加轻量的模型。如图4所示，该模块将卷积层分为两部分：一部分是常规的卷积操作，但输出的特征图会有严格限制；另一部分是通过线性变换生成，这可以减少计算量去生成更多信息。常规的卷积操作是卷积、批归一化BN和非线性激活组成，而线性变换是指普通卷积，不含批归一化和非线性激活。

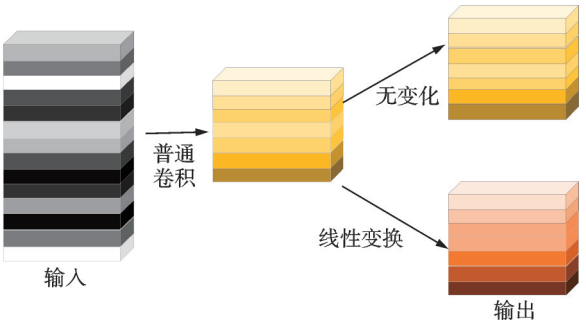


图4 Ghost模块[23]

Fig.4 Ghost module[23]

Ghost-ResNet网络结构依旧由残差模块和降采样残差模块组成，其改进的地方是：把模型中所有卷积层都替换成了Ghost模块，使残差模块由两个堆叠的Ghost模块组成，具体结构如图(5)所示。第1个Ghost模块主要是为了增加输入的通道数，第2个Ghost模块是减少通道数使Shortcut路径与输出通道数相匹配，再将输入和输出进行拼接得到最终输出。

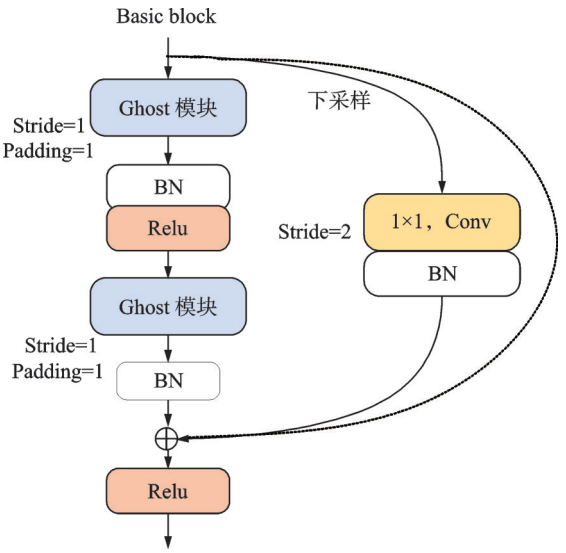


图5 残差模块

Fig.5 Residual module

对于之前的结构，对比Ghost-ResNet模型计算量，假设γ为d个$k\times k$卷积的计算开销，γ^*为d个Ghost模块的计算开销，则模型的计算量之比C_r的计算公式[23]为

$$C_r=d\frac{\gamma}{\gamma^*}=\frac{d\cdot n\cdot h'\cdot w''\cdot c\cdot k\cdot k}{\frac{n}{s}\cdot h'\cdot w'\cdot k\cdot k+(s-1)\cdot\frac{n}{s}\cdot h''\cdot w''\cdot a\cdot a}\approx$$
$$\frac{d\cdot c\cdot k\cdot k}{\frac{1}{s}\cdot c\cdot k\cdot k+\frac{s-1}{s}\cdot a\cdot a}\approx\frac{d\cdot s\cdot c}{s+c-1}\approx ds \tag{1}$$

这里输入张量为$c\cdot h'\cdot w'$，经过卷积后为$c\cdot h''\cdot w''$，卷积核大小为k，线性变换大小为d，s是指经过s变换，$\frac{n}{s}$是指第一次变换时输出通道数。由式(1)可知，原本的ResNet使用大量的普通卷积，会存在冗余特征，改后利用Ghost模块能较好地减少计算量。

为了保证在纹理信息提取部分加入的Ghost模块不影响其精度，在ResNet34-PSPNet的每个编码解码层的连接阶段加入能够收集通道空间信息的轻量型卷积注意力模块(CBAM)，使编码阶段的浅层特征与解码阶段的深层特征进行融合，该模块沿着通道和空间依次推断注意力图，能够进行自适应特征优化。

CBAM模块是一种结合空间与通道的注意力机制模块，能够同时兼顾通道和空间两个方面，获得更为丰富的特征图。如图6所示，其CBAM模块的结构在通道注意力模块部分，输入特征F经过最大

池化和平均池化用来聚合特征映射的空间信息,再分别送入多重感知机MLP,得到的两个输出特征进行加和操作再通过Sigmoid激活后生成通道注意力特征图$M_c(F)$。通道注意力子模块的计算过程如下

$$M_c(F)=\sigma\left(\mathrm{MLP}\left(\mathrm{AvgPool}(F)\right)+\mathrm{MLP}\left(\mathrm{MaxPool}(F)\right)\right)=\sigma\left(W_1(W_0(F^c_{\mathrm{avg}}))+W_1\left(W_0(F^c_{\mathrm{max}})\right)\right) \tag{2}$$

将其与F进行乘法操作得到空间注意力子模块的输入F'

$$F'=M_c(F)\odot F \tag{3}$$

在空间注意力模块中压缩通道,将前面得到的通道注意力特征作为输入,将基于通道的最大池化和平均池化做一个连接,通过7×7卷积进行降维,即$H\times W\times 1$,如式(4)所示。再经过Sigmoid生成空间注意力特征$M_s(F')$,将得到的两个模块的特征做乘法得到最终特征图F''。将CBAM模块插入到每个编码解码层的连接阶段,不仅连接了ResNet,得到丰富的特征图,还连接了PSPNet,得到更多上下文信息,增强有用特征,提高网络模型特征提取能力。

$$M_s(F')=\sigma\left(f^{7\times 7}\left(\left[\mathrm{AvgPool}(F');\ \mathrm{MaxPool}(F')\right]\right)\right)=\sigma\left(f^{7\times 7}\left(\left[F'^s_{\mathrm{avg}};\ F'^s_{\mathrm{max}}\right]\right)\right) \tag{4}$$

$$F''=M_s(F')\odot F' \tag{5}$$

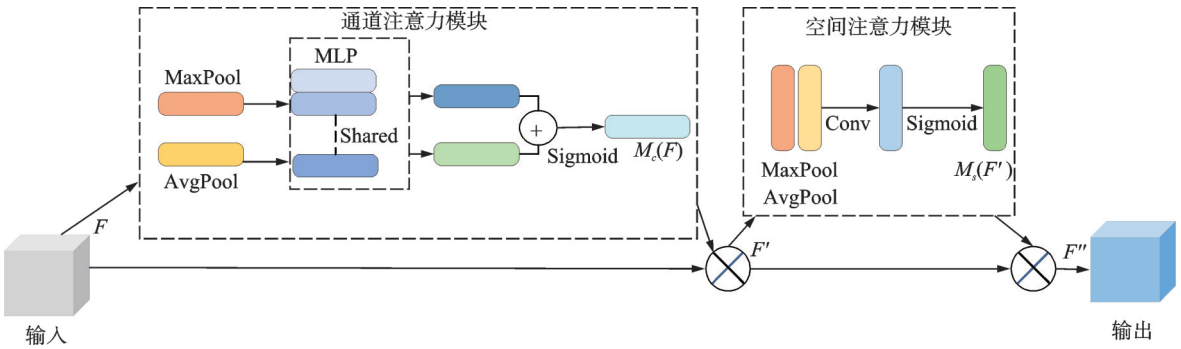


图6 CBAM模块网络结构图[15]

Fig.6 CBAM module network structure diagram[15]

1.2.2 深度特征提取网络

深度特征提取网络使用的是RandLA-Net网络,它是一个轻量级网络,采用此网络框架不仅计算效率高而且内存占用少,主要由随机采样和局部特征聚合模块组成。该网络之前用于三维点云的语义分割,采用的随机降采样能大大减少计算量,同时局部特征聚合模块用来解决随机降采样带来信息丢失的问题,但局部特征聚合只能在一定的领域内解决丢失问题,为了学习更多的几何信息,在相同数量的神经网络参数下,通过扩大KNN邻域来增加感受野。若仅仅用最后一层的特征,会导致缺少不同大小的特征,因此设计了一个多级分层特征融合网络,可以融合多尺度信息,获得更好的性能。基于RandLA-Net的主干网络也是编码-解码结构,首先扩展局部特征聚合[24],并和随机采样(Random sampling, RS)层组成4个编码层用来学习每个采样点的特征,然后对每一级下采样的特征进行上采样,并连接属于同一层的特征,最后使用3个全连接层和1个DP层来预测每个点的语义标签。

原本局部特征聚合模块LFA可分为3个模块:局部编码模块、注意力池化模块和扩张残差块。现提出的DLFA是将局部编码模块进行扩展,如图7所示,主要是为了学习更大邻域的局部特征,并利用自注意力池进行聚合特征。扩展局部编码模块是将KNN算法为每个点p_i找到最近的K个领域点变为了$2K$个点,然后又利用随机采样选K个点作为领域点进行后续计算,使得每个采样点扩大了2倍感受野。对得到的$\boldsymbol{p}_i$的K个最近邻点$\left\{\boldsymbol{p}_i^1\cdots\boldsymbol{p}_i^k\cdots\boldsymbol{p}_i^K\right\}$进行相对位置编码,将中心点三维坐标$\boldsymbol{p}_i$、领域点三维坐

标p_i^k、相对坐标$p_i-p_i^k$和欧氏距离$\|p_i-p_i^k\|$连接起来构成了l_i^k，如式(6)所示，其实这部分对学习邻域特征的输入l_i^k与原先的输入并没有改变，最终与邻域点对应的点特征f_i^k连接得到新的点特征$\hat{f}_i^k$，扩充的点特征集合如式(7)所示。

$$l_i^k=\mathrm{MLP}\left(p_i\oplus p_i^k\oplus\left(p_i-p_i^k\right)\oplus\|p_i-p_i^k\|\right) \tag{6}$$

$$\hat{F}_i=\left\{\hat{f}_i^1\cdots\hat{f}_i^k\cdots\hat{f}_i^K\right\} \tag{7}$$

将编码所得的拼接特征进行一个自注意力的学习，这是注意力池化模块，它不仅可以用来聚合先前融合的特征，还可以解决在空间上点云的无序性问题，这里用该模块代替了Max Pooling等对称函数，进一步对特征进行学习，对点云分割更加有利。点云数据会持续降采样，因此要扩大每个点的接受场，将两个扩展局部编码模块和注意力池化模块组成扩张残差块，以便尽可能保留输入点云的几何细节。提取完点云局部特征后，为了防止过度拟合，加上原始点云的特征，进行全局特征融合来优化网络，最终得到有效的几何特征。基于RandLA-Net网络在编码层和解码层只进行了简单的同尺度和同维度的跳跃连接，而这里在编码层和解码层进行拼接时，实现一个多级别的拼接，进行跨尺度的特征融合。多层次特征融合MHFF如图8所示，编码层进行下采样，解码层进行上采样，增加了不同层次上采样特征的结果(绿色部分)，对于大小相同的同一级别特征，采用特征拼接的方法进行集成。这样可以减少下采样和上采样丢失的一些特征，多层次特征融合结构的不同之处是增加了不同级别的上采样特征的结果，可以让相同大小的同一级别的要素进行连接集成，通过这种多层次融合，能够得到更加丰富的几何信息。

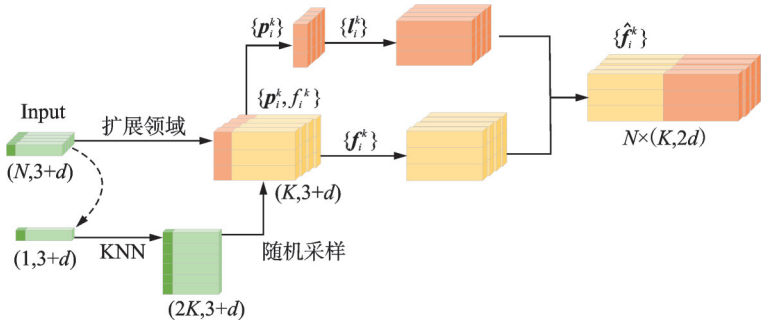


图7 扩展局部编码模块

Fig.7 Extended local coding module

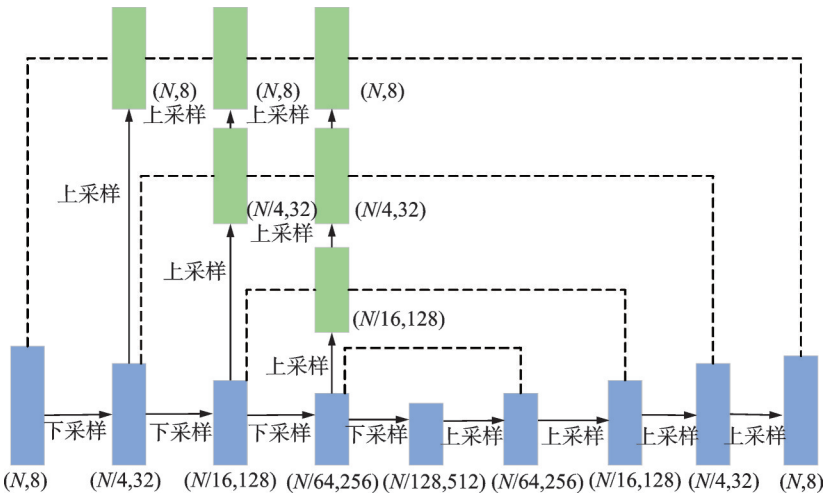


图8 多层次特征融合

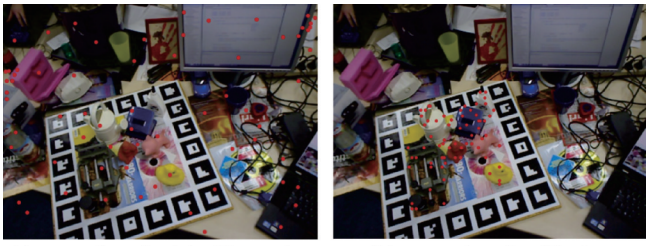
Fig.8 Multi-level hierarchical feature fusion

1.3 基于3D关键点的位姿估计

输出阶段由实例语义分割模块和3D关键点检测模块两部分组成。将特征提取模块获得逐点特征送至实例语义分割模块区分不同的对象实例，送到3D关键点检测模块用3D关键点投票方式来检测出

每个物体的3D关键点。

由于背景点对3D关键点检测没有帮助,反而会对其检测造成不便,这里利用实例分割出来的前景掩码通过滤波背景算法来过滤背景点,减少背景点的干扰,能大大提高算法效率,之后再进行位姿估计。为了得到有效的关键点,直接采用的3D关键点选择算法SIFT-FPS,采用SIFT算法检测纹理图像中特殊的2D关键点,再将其扩展到3D,这样SIFT算法检测的纹理信息与FPS算法[25]检测的欧氏距离相结合,能够充分利用对象纹理和几何信息,最后FPS算法能得到N个关键点,并确保被选关键点均匀分布在对象表面,更易于检测。原始关键点与过滤背景点的关键点如图9所示,可以看出过滤背景点的关键点选择更加准确。



(a) Extraction of original key points (b) Extraction after filtering background points

图9 选定关键点的结果对比

Fig.9 Comparison of results for selected key points

对于每个目标物体,3D关键点检测模块用来检测每个目标的关键点,这部分能够预测可见点到目标关键点的平移偏移量,再通过这些可见的坐标和预测的偏移量来得出目标关键点的位置作为投票点,投票点由聚类算法进行聚类以消除离群点干扰,群集的中心点被选为投票的关键点。给定一组可见点$\left\{p_i\right\}_{i=1}^{N}$和一组选定关键点$\left\{k_{p_j}\right\}_{j=1}^{M}$,$f_i^j$表示从第$i$个种子点到第$j$个关键点的平移偏移量,$f_i^{j^*}$指真实的平移偏移量,使用$L_1$损失函数来监督学习

$$L_{\text{keypoint}}=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\left\| f_i^j-f_i^{j^*}\right\|\Pi\left(p_i\epsilon I\right) \tag{8}$$

实例语义分割包含预测每点的语义标签的语义分割和学习每点到对象中心偏移量的中心点投票,以区分不同的实例。语义分割模块和中心投票模块都由Shared MLP组成。语义分割模块与3D关键点检测模块联合优化,通过提取实例的全局和局部特征来区分不同目标物体,有助于定位目标物体上一点,从而推出关键点偏移,也有助于区分外观相似大小不同的物体,损失函数采用Focal loss

$$L_{\text{semantic}}=-\alpha\left(1-q_i\right)^{\gamma}\log q_i \quad q_i=C_i\cdot I_i \tag{9}$$

式中:C为点云置信度,I为语义标签真值。中心点投票模块根据逐点特征来预测到目标对象中心的欧几里得平移偏移,$\Delta\boldsymbol{X}_i$是指所属目标对象中心的欧几里得平移偏移,$\Delta\boldsymbol{X}_i^*$指真实的平移偏移,使用L_1 loss进行监督学习

$$L_{\text{center}}=\frac{1}{N}\sum_{i=1}^{N}\left\|\Delta\boldsymbol{X}_i-\Delta\boldsymbol{X}_i^*\right\|\Pi\left(p_i\epsilon I\right) \tag{10}$$

3个模块联合优化,多任务学习损失函数:$L_{\text{multi-task}}=\lambda_1 L_{\text{keypoint}}+\lambda_2 L_{\text{semantic}}+\lambda_3 L_{\text{center}}$,其中$\lambda_1$、$\lambda_2$、$\lambda_3$分别为每个任务的权重。

给定两个点集,一个是来自物体坐标系中选定的3D关键点集$\left\{kp_j\right\}_{j=1}^{M}$,另一个是相机坐标系中与之对应的3D关键点集$\left\{kp_j'\right\}_{j=1}^{M}$,位姿估计模块是通过最小二乘拟合算法来计算$R$、$T$,如式(11)通过最小化平方损失来计算位姿估计$R$、$T$。

$$L_{\text{least-squares}}=\sum_{j=1}^{M}\left\|kp_j-\left(R\cdot kp_j'+T\right)\right\|^2 \tag{11}$$

2 实验和分析

2.1 数据集

在3个公共基准数据集上进行评估:LINEMOD、Occlusion LINEMOD和YCB-Video数据集。

LINEMOD数据集是目前较流行的单对象6D姿态估计数据集之一,它是由13个低纹理的家庭物品组成,这里每个物品都是在标有Aruco码的杂乱场景中标注了六自由度位姿,每个物品都有约1 200组的数据,包含了RGB图像和深度图、实例掩膜,图像大小为480 pixel×640 pixel。这类数据集常常用于场景杂乱、光照变化、低纹理物体。

Occlusion LINEMOD数据集是从LineMOD数据集中选择并在1 214张遮挡的图像中重新注释对象,每一帧都提供了多个注释对象用于测试。这类数据集具有严重遮挡、局部视图和照明条件变化的挑战。

YCB-Video数据集是在YCB数据集中选择了21个形状纹理各不相同的目标对象,这里通过RGBD相机获取的92个视频,每个视频里都含有在不同环境下3到9个目标对象,每个物体的数据都有标注的位姿信息、实例掩膜和三维模型。这类数据集同样具有变化的光照、图像噪声和严重遮挡问题等挑战

2.2 评价标准

对LINEMOD、Occlusion LINEMOD和YCB-Video数据集位姿估计时常用这两种评价指标对算法性能进行验证,分别是针对非对称物体的ADD度量和针对对称物体的ADD-S度量。ADD是指预测值RT和真值RT分别对模型点云进行变换后每个对应点距离的平均偏差,如式(12)所示。ADD-S考虑到对称物体,预测值和标注的真值在旋转轴上相差较大,使用距离最近的点来计算他们的平均距离偏差,如式(13)所示。

$$\mathrm{ADD}=\frac{1}{m}\sum_{x\in\mathrm{model}}\left\|(Rx+T)-\left(R_p x+T_p\right)\right\| \tag{12}$$

$$\mathrm{ADD\text{-}S}=\frac{1}{m}\sum_{x\in\mathrm{model}}\min_{x_2\in\mathrm{model}}\left\|(Rx_1+T)-\left(R_p x_2+T_p\right)\right\| \tag{13}$$

式中:x表示物体模型中的体素点,m表示点的总数,R、T表示预测的位姿,R_p、T_p表示位姿的真值。

测评时用ADD来度量上述数据集中的非对称物体,对称物体用ADD-S度量,根据上述的评估公式,若计算平均距离误差小于物体模型直径的10%,则其预测的位姿估计是正确的。

2.3 实验细节

对深度学习网络模型和训练模型是基于Pytorch1.12和python3.8环境下的,在RTX3050 GPU硬件平台上完成的,训练了20个epochs,使用优化器是Adam,学习率设为0.001。纹理特征提取中下采样部分用Ghost-ResNet的4个Ghost模块和PSPNet金字塔池化模快,解码部分用PSPNet进行上采样;深度信息提取中,先对深度图随机采样12 288个点经过全连接层处理后,送到RandLA-Net的4个编码层进行随机采样和局部特征聚合,再采用上采样和MLP进行解码。将编码和解码每层都用Share MLP进行两种特征逐点融合,在融合8次后得到双向融合的密集特征,再通过MLPs组成实例语义分割和3D关键点投票模块。对于关键点,应用的SIFT-FPS算法为每个物体选择了8个关键点。

2.4 实验结果与评价分析

2.4.1 LINEMOD数据集实验结果与分析

在基于LINEMOD数据集的实验中,将本文提出的方法与PoseCNN[26]、DenseFusion、BiCo-Net以及FFB6D方法进行对比,基于RGB的PoseCNN方法其因缺失深度信息准确度低但速度快,基于

RGB-D的DenseFusion、BiCo-Net以及FFB6D方法精度较高,但速度慢实时性差。对比以上4种方法,本文基于改进的FFB6D的方法效果最好,在速度上提升了30%,其精度ADD(S)指标平均值从99.7%提升到了99.8%,如图10所示是在LINEMOD数据集上给出了可视化的位姿检测结果,绿色是真实情况,红色是姿态估计算法预测的3D框,可以看出在该算法下红色3D框与绿色3D框紧密重合。

图10 在LINEMOD数据集上位姿检测结果

Fig.10 Position detection results on LINEMOD dataset

表1对比了PoseCNN、DenseFusion、BiCo-Net及FFB6D这4种方法,并列出了LINEMOD数据集中每个类别的ADD(S)准确率。通过实验数据可以发现本文所提的姿态估计精度高于现有同类的方法,即使没有细化过程,也达到了99.8%。在多个类别中取得了最高的分数,包括Bench vise、Camera、Cat、Lamp、Hole puncher,验证了该方法具有更出色的性能。表2为本文方法在Occlusion LINEMOD数据集上与PoseCNN、PVN3D 、BiCo-Net以及FFB6D的比较结果,其中PVN3D、FFB6D是基于3D关键点的算法,它能有效用于遮挡环境下,而BiCo-Net引入了点姿态敏感正则化,有效地提高了局部特

表1 在LINEMOD数据集上不同算法的结果对比

Table 1 Comparison of results of different algorithms on LINEMOD dataset %

Object	PoseCNN	DenseFusion	FFB6D	BiCo-Net	Ours
Ape	77.0	92.3	99.1	97.3	99.4
Bench vise	97.5	93.2	100.0	98.8	100.0
Camera	93.5	94.4	100.0	99.6	100.0
Can	96.5	93.1	100.0	99.3	100.0
Cat	82.1	96.5	99.5	100.0	99.8
Driller	95.0	87.0	99.7	98.9	99.9
Duck	77.7	92.3	98.4	98.7	98.6
Eggbox	97.1	99.8	99.5	99.8	99.7
Glue	99.4	100.0	100.0	99.8	100.0
Hole puncher	52.8	92.1	100.0	99.2	99.9
Iron	98.3	97.0	99.9	100.0	99.8
Lamp	97.5	95.3	99.8	99.7	99.9
Phone	87.7	92.8	99.6	99.2	99.7
Mean	88.6	94.3	99.7	99.3	99.8

征编码的判别能力，其ADD可达到69.5%，但可以发现本文方法还是比PoseCNN、PVN3D、FFB6D的准确度有所提高，且ADD达到66.3%，这是因为本文充分利用了关键点之间的几何关系，得到准确的位姿估计。

表2 在Occlusion LINEMOD数据集上不同算法的结果对比

Table 2 Comparison of results of different algorithms on Occlusion LINEMOD dataset

%

Method	ADD
PoseCNN	24.9
PVN3D	62.5
FFB6D	64.4
BiCo-Net	69.5
Ours	66.3

2.4.2 YCB-Video数据集实验结果与分析

在基于YCB-Video数据集实验中，与PoseCNN、DenseFusion、ICG和FFB6D方法进行比较，同样使用ADD-S和ADD(S)两种评价标准，得到以下21个类别在5个不同方法的测试结果如表3所示，同样本文方法性能高于其他4种，可以看出有效融合纹理和深度信息能大大提升性能。由于YCB-Video数据集常用于挑战遮挡环境下，FFB6D方法会明显优于DenseFusion方法，ICG在ADD-S度量方面达到了不错结果，但ADD度量效果较差，这是由于无纹理方法使某些非对称物体的不确定性，ICG+在其基础上有所改善，但它是一种3D物体跟踪方法，需要高昂的计算成本。本文方法与FFB6D一样都是基于关键点的特征融合算法，可以在严重遮挡下利用关键点之间的几何关系使位姿估计的精度更加准确，其ADD-S和ADD(S)分别达到96.7%和94%。图11展示了在YCB-Video数据集上不同方法直观对比结果，图中的点是三维点云物体模型通过估计的姿态参数和相机内参数作用，映射到图片后的结果，不同颜色的点代表了来自不同物体的点，用红框标注识别到的物体，如第1列

表3 在YCB-Video数据集上不同算法的结果对比

Table 3 Comparison of results of different algorithms on YCB-Video dataset

%

Object	PoseCNN		DenseFusion		FFB6D		ICG		Ours	
	ADD-S	ADD(S)	ADD-S	ADD(S)	ADD-S	ADD(S)	ADD-S	ADD(S)	ADD-S	ADD(S)
002_Master_chef_can	83.9	50.2	95.3	70.7	96.4	80.7	89.7	66.4	95.7	81.1
003_cracker_box	76.9	53.1	92.5	86.9	96.4	95.0	92.1	82.4	96.6	95.3
004_sugar_box	84.2	68.4	95.1	90.8	97.7	96.8	98.4	96.1	98.0	97.1
005_tomato_soup_can	81.0	66.2	93.8	84.7	95.8	88.1	97.3	73.2	97.3	89.4
006_mustard_bottle	90.4	81.0	95.8	90.9	98.1	97.6	98.4	96.2	97.7	97.4
007_ tuna_fish_can	88.0	70.7	95.7	79.6	97.2	91.3	95.8	73.2	96.9	91.6
008_pudding_box	79.1	62.7	94.3	89.3	96.3	93.1	88.9	73.8	97.2	94.1
009_gelatin_box	87.2	75.2	97.2	95.8	97.8	95.8	98.8	97.2	97.7	95.7
010_potted_meat_can	78.5	59.5	89.3	79.6	92.6	89.8	97.3	93.3	91.2	89.7
011_banana	86.0	72.3	90.0	76.7	97.4	94.9	98.4	95.6	97.6	94.7
019_gelatin_box	77.0	53.3	93.6	87.1	97.7	97.0	98.8	97.0	95.9	95.3
021_bleach_cleanser	71.6	50.3	94.4	87.5	96.5	93.7	97.5	92.6	97.3	93.7
024_bowl	69.6	69.6	86.0	86.0	95.8	95.8	98.4	74.4	97.0	97.0
025_mug	78.2	58.5	95.3	83.8	97.5	95.3	98.5	95.6	97.7	95.1
035_power_drill	72.7	55.3	92.1	83.7	97.3	96.2	98.5	96.7	96.9	96.0
036_wood_block	64.3	64.3	89.5	89.5	93.1	93.1	97.2	93.5	93.3	93.3
037_scissors	56.9	35.8	90.1	77.4	98.1	97.1	97.3	93.5	97.8	96.4
040_lager_marker	71.7	58.3	95.1	89.1	96.9	90.0	97.8	88.5	97.5	90.1
051_large_clamp	50.2	50.2	71.5	71.5	96.8	96.8	96.9	91.8	96.6	96.6
052_extra_large_clamp	44.1	44.1	70.2	70.2	96.1	96.1	94.3	85.9	96.9	96.9
061_foam_brick	88.0	88.0	92.2	92.2	97.6	97.6	98.5	96.2	98.0	98.0
MEAN	75.8	59.9	91.2	82.9	96.6	92.7	96.5	86.4	96.7	94.0

对香蕉的估计和第3列对碗的估计看出本文方法点云分布更紧密贴切目标物体,第2列中其他方法对剪刀都给出了相差较大的估计,而本文方法给出了更符合事实的预测。

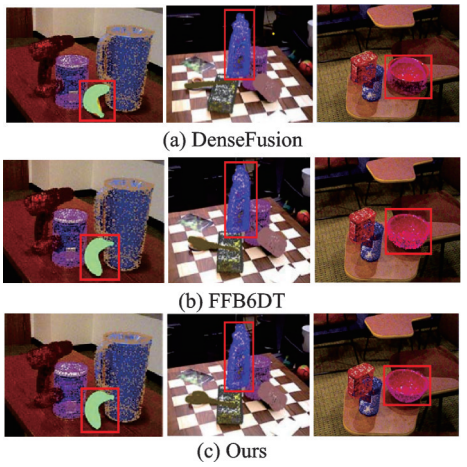


图11 在YCB-Video数据集中位姿检测结果

Fig.11 Position detection results on YCB-Videa dataset

2.4.3 消融实验和对比实验

本文方法是基于RGB-D的姿态估计,有两个主干网络分别基于RGB图像的Ghost-ResNet和基于点云改进的RandLA-Net,对它们进行消融实验和算法对比实验。

在表4中,本文使用参数量Params、浮点运算(Floating point operations, FLOPs)、ADD(S)等指标,在YCB-Video数据集上评估网络的有效性。首先,对RGB图像部分,以ResNet34为基础,加入Ghost模块后的模型是否减少了计算量,结果显示本文设计的网络结构比原先框架的ResNet网络减少了50%的参数量,参数量变小了,其计算量也减少到原来的一半,说明加入Ghost模块后能够使网络轻量化。在网络参数量减少的同时,为保证准确度不下降,对加入的CBAM注意力机制进行测试,可以看出在使用CBAM注意力机制后,参数量仍比原网络减少,且提高了精度。

表4 不同算法的参数量和速率对比

Table 4 Comparison of parameter counts and rates for different algorithms

模型	Params/10^6	FLOPs/B	ADD(S)/%
ResNet34	21.8	4.0	88.0
Ghost-ResNet34	9.0	2.1	87.7
Ghost-ResNet34-CBAM	11.3	2.4	91.2

在深度图部分,本文在原来基础上扩展了局部特征聚合并进行多层次特征融合,为了验证改进部分对模型的影响,将对深度信息提取网络进行消融实验,得到表5的结果,依旧使用了两种度量即参数量和ADD(S),可以看出DLFA由于扩大了提取特征的感受野,所以能比之前的模块提取更全面的几何信息,加入MHFF使性能得到了提高,这说明多层次融合是有必要的,能够利用低层次的特征来补充物体的信息。

表5 点云网络的消融实验

Table 5 Ablation experiments on point cloud networks

模型	Params/10^6	ADD(S)/%
LFA	23.3	92.7
DLFA	23.5	93.1
DLFA+MHFF	23.5	94.0

最后,为了评估背景点过滤后的实验结果,与之前使用的所有种子点的实验相比,这里用FFB6D算法与本文所提方法进行运行速度方面的比较,结果如表6所示。不管是本文方法还是FFB6D加入滤波背景算法(FBP)都比不加的效果要好很多,本文的方法运行速度相比先前提高了30%,且所需参数量更少。这是因为用所有种子点来计算背景点和边缘点与中心点的距离会浪费大量的时间,而本文方法通过语义分割获得的前景掩膜用来过滤掉背景点,这样使得最小二乘拟合阶段只需计算较少数量前景点的最终6D姿态参数,所以会使得计算速度很快。

表6 有无背景点过滤处理的检测速率对比

Table 6 Comparison of detection rates with and without background point filtering treatment

Method	FBP	Params/10^6	Runtime/(ms·frame^{-1})
FFB6D		33.8	142
FFB6D	√	33.4	100
Ours		23.5	137
Ours	√	23.2	93

3 结束语

针对在复杂环境、设备资源有限的情况下,利用深度相机如何获得目标物体准确的位姿信息,提出了一种基于双向融合纹理和深度信息的目标物体位姿检测方法。为了减少模型的计算量和参数量,将Ghost模块引入ResNet网络,解决了冗余特征问题,使网络的计算量变小,更好地部署在嵌入式设备上,并使用CBAM模块采集通道空间信息,提高了模型对特征信息的敏感度,有效提取纹理特征保证其精度;在点云网络设计了基于RandLA-Net的扩展局部特征聚合,能够增加邻域的感受野,弥补点云经下采样之后特征信息丢失,通过多级分层特征融合网络,融合多尺度信息,兼顾全局特征和局部特征,获得丰富的几何特征。最终在LINEMOD、Occlusion LINEMOD和YCB-Video公共数据集上得到了验证,本文改进的方法其准确度分别达到了99.8%、66.3%和94%,同时利用实例语义分割中获得的前景掩码来过滤背景点,提高了关键点选择的效率,速度也比先前快了1.3倍,加快了模型的训练速度。

参考文献:

[1] ZENG A, SONG S, YU K T, et al. Robotic pick and place of novel objects in clutter with multi-affordance grasping and cross-domain image matching[C]//Proceedings of 2018 IEEE International Conference on Robotics and Automation (ICRA). Brisbane: IEEE, 2018: 3750-3757.

[2] 王太勇,孙浩文.基于关键点特征融合的六自由度位姿估计方法[J].天津大学学报,2022,55(5): 543-551.
WANG Taiyong, SUN Haowen. Six degrees of freedom pose estimation based on keypoints feature fusion[J]. Journal of Tianjin University, 2022, 55(5): 543-551.

[3] 陈海永,李龙腾,陈鹏,等.复杂场景点云数据的6D位姿估计深度学习网络[J].电子与信息学报,2022,44(5): 1591-1601.
CHEN Haiyong, LI Longteng, CHEN Peng, et al. 6D pose estimation network in complex point cloud scenes[J]. Journal of Electronics & Information Technology, 2022, 44(5): 1591-1601.

[4] 孙晴艺.基于RGB-D数据的目标物体6D位姿估计研究[D].南宁:广西大学, 2022.
SUN Qingyi. Research on object 6D pose estimation based on RGB-D data[D]. Nanning: Guangxi University, 2022.

[5] 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]//Proceedings of Asian Conference on Computer Vision. Berlin: Springer, 2012: 548-562.

[6] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[7] RUSU R B, BLODOW N, BEETZ M. Fast point feature histograms (FPFH) for 3D registration[C]//Proceedings of 2009 IEEE International Conference on Robotics and Automation. [S.l.]: IEEE, 2009: 3212-3217.

[8] TOMBARI F, SALTI S, STEFANO L D. Unique signatures of histograms for local surface description[C]//Proceedings of European Conference on Computer Vision. Berlin: Springer, 2010: 356-369.

[9] KEHL W, MANHARDT F, TOMBARI F, et al. SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again[C]//Proceedings of the IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2017: 1521-1529.

[10] TEKIN B, SINHA S N, FUA P. Real-time seamless single shot 6D object pose prediction[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 292-301.

[11] PENG S, LIU Y, HUANG Q, et al. PVNet: Pixel-wise voting network for 6DoF pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 4561-4570.
[12] WANG C, XU D, ZHU Y, et al. DenseFusion: 6D object pose estimation by iterative dense fusion[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2019: 3343-3352.
[13] HE Y, SUN W, HUANG H, et al. PVN3D: A deep point-wise 3D keypoints voting network for 6D of pose estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2020: 11632-11641.
[14] HE Y, HUANG H, FAN H, et al. FFB6D: A full flow bidirectional fusion network for 6D pose estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2021.
[15] STOIBER M, SUNDERMEYER M, TRIEBEL R. Iterative corresponding geometry: Fusing region and depth for highly efficient 3D tracking of textureless objects[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans: IEEE, 2022; 6845-6855.
[16] STOIBER M, ELSAYED M, REICHERT A E, et al. Fusing visual appearance and geometry for multi-modality 6DoF object tracking[C]//2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). [S.l.]: IEEE, 2023: 1170-1177.
[17] XU Z L, ZHANG Y C, CHEN K, et al. BiCo-Net: Regress globally, match locally for robust 6D pose estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2022.
[18] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2016: 770-778.
[19] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]//Proceedings of the 2018 European Conference on Computer Vision. Cham: Springer, 2018: 3-19.
[20] HU Q, YANG B, XIE L, et al. EandLA-Net: Efficient semantic segmentation of large-scale point clouds[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2020: 11108-11117.
[21] DE VRIES A P, MAMOULIS N, NES N, et al. Efficient KNN search on vertically decomposed data[C]//Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data. [S.l.]:ACM, 2002: 322-333.
[22] ZHAO H S, SHI J P, QI X J, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017: 2881-2890.
[23] HAN K, WANG Y H, TIAN Q, et al. GhostNet: More features from cheap operations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2020: 1580-1589.
[24] FAN X Y, WANG L, JIANG S, et al. Dilated nearest-neighbor encoding for 3D semantic segmentation of point clouds[C]//Proceedings of the 2021 IEEE International Conference on Real-Time Computing and Robotics. [S.l.]: IEEE, 2021.
[25] MOENNING C, DODGSON N A. Fast marching farthest point sampling for point clouds and implicit surfaces[R]. Cambridge: University of Cambridge, 2003.
[26] XIANG Y, SCHMIDT T, NARAYANAN V, et al. PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes[EB/OL]. (2017-11-01).https://doi.org/10.48550/arXiv.1711.00199.

作者简介:

张亚炜(1999-),女,硕士研究生,研究方向:计算机视觉, E-mail: yaweiyaye@163.com。

付东翔(1971-),**通信作者**,男,副教授,研究生导师,研究方向:光电测试与智能系统、计算机视觉,E-mail: fudx@usst.edu.cn。

(编辑:夏道家)