基于伪孪生神经网络的低纹理工业零件6D位姿估计

王神龙,雍 宇,吴晨睿
(上海理工大学机械工程学院,上海 200093)

摘 要: 从单帧RGB图像中获取目标物体的6D位姿信息在机器人抓取、虚拟现实、自动驾驶等领域应用广泛.本文针对低纹理物体位姿估计精度不足的问题,提出一种基于伪孪生神经网络的位姿估计方法.首先,通过渲染CAD模型的方式,获取不同观察角度下的RGB图作为训练样本,解决了深度学习中数据集获取与标注较为繁琐的问题.其次,利用伪孪生神经网络结构学习二维图像特征和物体的三维网格模型特征之间的相似性,即分别采用全卷积网络和三维点云语义分割网络构成伪孪生神经网络,提取二维图像和三维模型的高维深层特征,使用网络推断密集的二维-三维对应关系.最后,通过PnP-RANSAC方法恢复物体的位姿.仿真数据集的实验结果表明,本文提出的方法具有较高的准确性和鲁棒性.

关键词: 深度学习;6D位姿估计;仿真数据集;伪孪生神经网络;点向密集匹配
基金项目: 国家自然科学基金青年项目(No.52105525);国家自然科学基金面上项目(No.12172226)
中图分类号: TP391.41 **文献标识码:** A **文章编号:** 0372-2112(2023)01-0192-10
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20211688

6D Pose Estimation of Low Texture Industrial Parts Based on Pseudo-Siamese Neural Network

WANG Shen-long, YONG Yu, WU Chen-rui
(*College of Mechanical Engineering*, *University of Shanghai for Science and Technology*, *Shanghai* 200093, *China*)

Abstract: Obtaining the 6D pose information of the target object from a single frame RGB image is widely used in the fields of robot capture, virtual reality, automatic driving, and so on. Aiming at the problem of insufficient accuracy of pose estimation of low texture objects, a pose estimation method based on pseudo-siamese neural network is proposed in this paper. Firstly, RGB images from different viewing angles are obtained as training samples by rendering CAD models, which solves the cumbersome problem of data set acquisition and annotation in deep learning. Secondly, the pseudo-siamese neural network structure is used to learn the similarity between the two-dimensional image features and the three-dimensional mesh model features of the object, that is, the full convolution network and the three-dimensional point cloud semantic segmentation network are used to form the pseudo-siamese neural network, extract the high-dimensional deep features of the two-dimensional image and the three-dimensional model, and use the network to infer the dense two-dimensional three-dimensional correspondence. Finally, the pose of the object is restored by PNP-RANSAC method. The experimental results of simulation data sets show that the proposed method has high accuracy and robustness.

Key words: deep learning; 6D pose estimation; simulation data set; pseudo-siamese neural network; dense matching of points
Foundation Item(s): National Natural Science of China (No.52105525, No.12172226)

1 引言

6D位姿估计是从物体坐标系到相机坐标系的刚性转换,包括旋转量R和平移量T[1],其在虚拟现实、机器人视觉、自动驾驶、增强现实等领域发挥着十分重要的作用.随着智能制造行业的快速发展,工业零件的位姿估计已经成为装配、堆垛和人机协同等应用场景的关键技术[2-4].

物体的信息可以通过RGB传感器和深度传感器获

收稿日期:2021-12-23;修回日期:2022-03-23;责任编辑:朱梅玉

得. 根据输入数据的不同,6D位姿估计方法可以分为三类,分别为基于RGB输入[5~13]、基于点云输入[14,15]和基于RGB-D[16~18]输入的方法. 其中,由于点云的稀疏性或者无序性,以点云为输入方式的位姿估计方法往往鲁棒性较差. 以RGB-D为输入的方式,结合了两种数据的特征,虽然位姿估计的鲁棒性得到保证,但是数据占用的计算空间过大,深度传感器设备较为昂贵且不易获得,限制了该方式的应用. 因此,本文主要关注较为低成本的基于RGB的位姿估计方法. 针对纹理特征明显的目标对象,传统方法利用图像特征描述符实现纹理对象的位姿估计[19],这些描述符能直观地利用图像梯度信息为对象进行分类. 二维与三维之间的对应关系可以通过这些描述符的相似性来确定,进而利用PnP(Perspective-n-Points)算法[20]求解目标位姿. 然而,许多低纹理的物体,如纸盒、金属零件和塑料零件,都存在于复杂的工业环境中,这些低纹理物体在图像中几乎没有显示出明显的颜色变化,旋转不变的描述符无法匹配其真正特征.

近些年,深度学习的发展提高了对低纹理或者无纹理物体进行位姿估计的技术. 基于RGB输入的位姿估计方法涌现出许多研究成果,根据恢复位姿方式的不同可以分为直接法、关键点法、坐标对应法等几类方法. 在文献[5~7]中的方法旨在直接回归物体的6D位姿. 例如,Xiang等[5]提出用深度神经网络(Deep Neural Networks,DNN)直接回归物体的旋转和平移;Kendall等[6]提出使用卷积神经网络(Convolutional Neural Networks,CNN)结构从单个RGB图像直接回归相机位姿;Kehl等[7]提出了一种基于分类离散视点的方法来预测真实姿态的粗略离散近似值. 上述方法被称为直接法,然而,受环境光以及旋转空间的非线性等因素的较大影响,这些方法的鲁棒性往往较差.

相较直接法,关键点法在空间中降低了回归目标的数据量,使得其恢复位姿的效率更快. 文献[8,9]旨在从RGB图中学习选取目标的关键点,并通过PnP回归物体的位姿. 例如,PVNET[8]用修改后的全卷积网络(Fully Convolutional Networks,FCN)对指向关键点的像素级向量进行回归,并使用这些向量来投票选择关键点位置,将得分最高的关键点通过PnP算法来恢复物体的位姿. DPVL[9]通过建立新的损失函数,在PVNET中根据投票方向、像素和关键点的基础上结合距离因素让网络的收敛速度和精度得到进一步的提高. 然而此类方法由于透视投影导致的几何信息丢失容易造成误匹配等问题.

与关键点方法不同,坐标对应法通过密集点对关系减少了选取关键点错误而出现误匹配的情况. 文献[10~12]中的方法都是通过不同方式去定义密集点对关系来恢复物体的6D位姿. CDPN[10]是第一种将目标的三维坐标系用于二维和三维之间的密集对应关系上的网络,虽然保持了较高的准确率,但缺少对具有对称性物体的考虑. DPOD[11]通过UV图谱构建二维和三维对应关系,根据给定像素颜色映射估计其在三维模型表面上的对应位置,从而提供图像像素与三维模型顶点之间的关系. PSGMN[12]分别对二维图像中的目标对象以及三维模型进行特征编码,随后利用位姿估计网络学习这两种编码特征的密集匹配关系来求解物体的位姿,以大量的点对应关系使姿态估计结果更加稳定和准确. 然而,低纹理工业零件存在较大的曲率变化和复杂的形状等问题,PSGMN中的SplineCNN子网络对一些小巧的目标具有很好的特征提取能力,但却丧失了对一些重要3D特征的提取. 针对上述各类问题,本文的创新点与工作主要有以下几点.

(1)虽然深度学习对无纹理物体的位姿估计问题得到解决,但是鉴于文献[8~12]等位姿估计方法都是将LINEMOD等开源数据集作为训练样本,其估计的目标对象并不适用于实际工业场景,所以本文通过渲染工业零件CAD模型获取不同观察角度下的RGB图、掩码图以及位姿信息作为训练样本,增强数据集的多样性. 这样不仅解决了深度学习中大规模工业零件数据集获取与标注较为繁琐的问题,而且通过生成的仿真数据集提高了6D位姿估计的准确性和模型泛化能力.

(2)一般的网络设计并不具有场景鲁棒性. 在工业场景中零件的规格是多变且不一的,它们都具有复杂的外观形状,SPlineCNN对于较大物体的特征提取也并不适用,为此本文基于伪孪生神经图匹配网络框架[12],通过使用RandLA-Net[21]代替SplineCNN来降低网络模型参数量,在降低计算成本的同时,新的网络仍具有很高的位姿识别精度.

为了更加清晰地展示本文的研究思路与研究内容,主要研究技术路线如图1所示.

2　仿真数据集的建立

现有的基于深度学习的6D位姿估计方法大多是在开源的LINEMOD,OCCLUSION_LINEMOD等数据集上进行测试. 然而,由于工业零件的特殊性,在这些数据集上测试效果很好的网络并不能适用于一些低纹理的、结构复杂的工业零件. 基于该问题,本文提出了一种用于工业零件位姿估计的仿真数据集生成方法.

2.1　低纹理工业零件建模

对图2(a)相机所拍摄的低纹理工业零件进行原尺寸的测绘建模,使用CATIA三维建模软件完成建模,效果如图2(b)所示,该工业零件用于5G天线装置. 将CAD模型导入MESHLAB进行点云化处理,经过蒙特卡

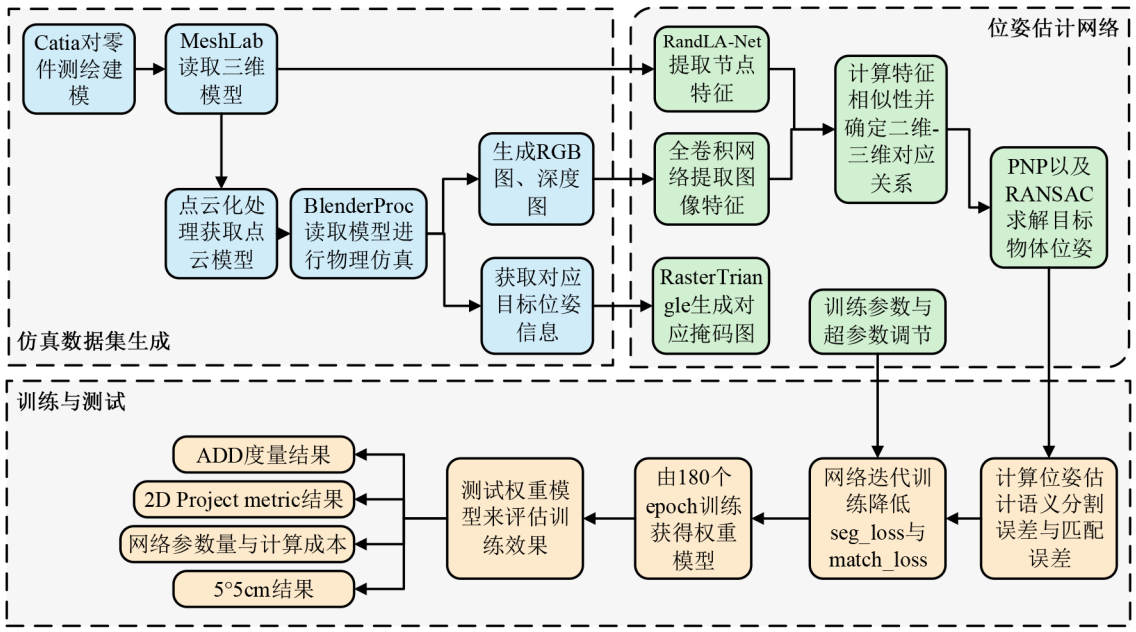
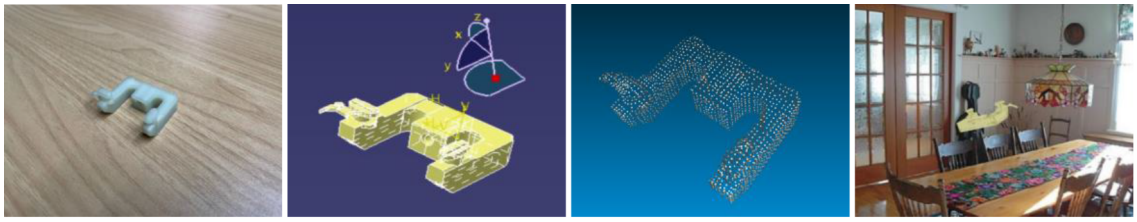


图1 本文研究技术路线图



(*a*) 零件实物图 (*b*) 三维建模图 (*c*) 点云模型图 (*d*) 仿真渲染图

图2 工业零件仿真数据集的建立

洛滤波、泊松采样等处理,获得采样顶点3 108个,三角面片6 692个,同时将零件实际RGB值赋给点云模型,图2(*c*)为MESHLAB点云化处理后的点云模型.

2.2 仿真数据集生成

首先,利用Raster triangle[16]生成仿真图数据集,渲染结果如图2(*d*)所示,零件位姿并非清晰可见,且该渲染程序并没有考虑到零件实际所处的真实环境,即重力大小、光照强度和方向等因素对工业零件产生的实际影响,而这些因素往往会直接影响位姿估计网络的训练和测试结果.

其次,本文利用改进的Blender proc[22]将建立好的3D模型预设置随机位姿,置于环境上方,如图3所示.零件所处环境由预先设定好的固定光照强度、光照方向、重力参数等来模拟真实环境,同时对零件本身预设好表面粗糙度、透光性等参数,接着零件依靠重力掉落,模型产生物理碰撞后静止.

最后,在模拟相机视野下渲染得到零件的RGB图,同时获取工件地面真实位姿与之后生成的掩码图对应.

改进之后的渲染效果如图4所示,针对不同的目标

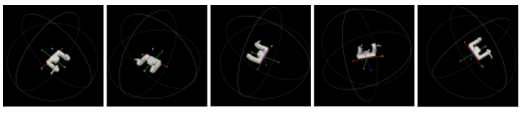


图3 目标的三维坐标轴以及虚拟观察采样点

对象,各生成8 000张仿真图片,其中7 000张应用于训练集,剩余1 000张用于测试集.

3 位姿估计网络

孪生神经网络[23]又称"连体神经网络",其"连体"通过共享权值来实现.若使用两个神经网络分别对图片进行特征提取,提取到的特征很可能不在一个分布域中,此时可以考虑使用一个神经网络进行特征提取再进行比较.因此,孪生神经网络可以提取出两个输入图片同一分布域的特征,从而判断两个输入图片的相似性.狭义的孪生神经网络由两个结构相同且权重共享的神经网络拼接而成.广义的孪生神经网络即"伪孪生神经网络",可由任意两个神经网络拼接而成.

本文将使用一种伪孪生神经网络[12]来评估关键点之间的相似性.

从单一RGB图像中估计对象位姿的核心思想是找

(a) 工业零件　(b) 黄鸭　(c) 相机　(d) 花洒　(e) T-Less 零件

图4　仿真数据集

到图像中对象与其三维模型之间的局部对应关系,如式(1)所示:

$$q=\boldsymbol{K}\left[R|T\right]Q \tag{1}$$

对象相机位姿坐标可以通过图像投影模型获得. 其中,

$$\boldsymbol{K}=\begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

q是物体三维模型上的三维点Q在图像上的二维投影;$\boldsymbol{K}$表示表示相机坐标系与图像坐标系之间转换的相机参数;R和T分别代表物体坐标相对于相机坐标的旋转量和平移量. 在本文的网络中,只要能准确得到至少3对二维-三维对应,就可以通过PnP算法求解出R和T. 因此,本文位姿估计网络的最重要的步骤是获得精确的二维-三维对应关系.

位姿估计网络结构如图5所示,主要由3个子模块构成,包括利用全卷积网络的图像特征提取模块、利用三维点云语义分割网络的节点特征提取模块和位姿估计模块.

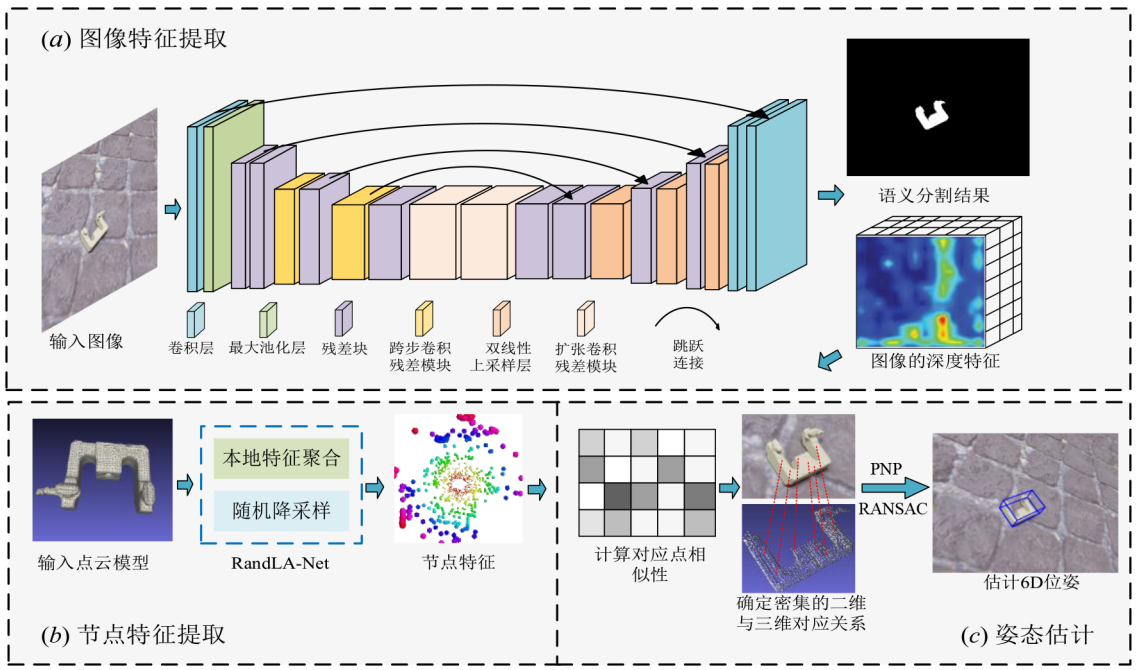


图5　位姿估计网络结构图

3.1　图像特征提取

在图像特征提取子模块中,本文利用进一步改进的全卷积网络模型来提取二维图像的深度特征. 如图5(a)所示,网络模型由数层组成. 利用残差块对输入特征进行卷积操作,利用最大池化层对输入特征进行下采样,进而从网络中提取出不同尺度的特征. 利用跳跃连接残差块的特征映射来保留目标图像的初始信息. 双线性上采样层用于获取目标对象的语义分割信息.

本文将全卷积网络模型的函数定义为$\vartheta_J=\Phi_{\theta_1}(J)$,其中,$J$表示输入图像,$\theta_1$表示模型$\Phi$的参数. 一般的全卷积网络模型输出具有$(C+1)\times H\times W$的形式,其中,$C$是对象的类别数,维度1表示背景,$H$和$W$分别表示二维图像的高度和宽度. 由于本文估计的目标对象仅为一个,所以对网络模型进一步修改,输出形式为$(D+1)\times H\times W$的张量,其中D表示二维图像中每个像素的D维局部特征,1表示目标对象所处虚拟环境背景. 因此,输出由两部分组成,其中$M_J=1\times H\times W$表示目标对象按像素分割的类概率,$H_J=D\times H\times W$表示每个像素的高维局部特征.

3.2 节点特征提取

本文利用了多层RandLA-Net[21]，一种高效的神经网络架构，使用简单而快速的随机采样方法来大大降低点密度，同时应用一个局部特征聚合器来保留显著的节点特征. 本文利用Python中的numpy来生成K个索引，利用这些指标从点云中收集相应的空间坐标和点特征，进而完成一次随机降采样. 在局部特征聚合部分，首先利用相对点位置编码搜索中心点p_i最近的K个点，即

$$\boldsymbol{r}_i^k=\mathrm{MLP}\left(p_i\oplus p_i^k\oplus\left(p_i-p_i^k\right)\oplus\left\|p_i-p_i^k\right\|\right) \tag{3}$$

其中，p_i和p_i^k为点的x-y-z位置，$\oplus$为连接操作，$\|\cdot\|$计算相邻点与中心点之间的欧氏距离. 对于每个相邻点p_i^k，将编码的相对点位置$\boldsymbol{r}_i^k$与其对应的点特征相连得到一个增广特征向量$\hat{\boldsymbol{f}}_i^K$. 最终，输出一组新的相邻特征$\hat{\boldsymbol{F}}_i=\left\{\hat{\boldsymbol{f}}_i^1\cdots\hat{\boldsymbol{f}}_i^k\cdots\hat{\boldsymbol{f}}_i^K\right\}$. 接下来，利用一个共享函数$g$来自动学习重要的局部特征，定义为

$$B_i^k=g\left(\hat{\boldsymbol{f}}_i^k,\boldsymbol{W}\right) \tag{4}$$

其中，$\boldsymbol{W}$为可学习权值. 接着再对学习到的注意力得分加权求和，即

$$\tilde{\boldsymbol{f}}_i=\sum_{k=1}^{K}\left(\hat{\boldsymbol{f}}_i^k\cdot B_i^k\right) \tag{5}$$

其中，$\tilde{\boldsymbol{f}}_i$为高维特征向量. 最后通过扩张的残差块有效地保留复杂的局部结构.

综上所述，本文采用四层RandLA-Net来构造网格特征提取网络$\Upsilon_{\theta_2}\left(X_n,\mathcal{E}_n,\boldsymbol{U}_n\right)$，其中，$\mathcal{E}_n$表示边集，$\boldsymbol{U}_n$表示所有节点的伪坐标矩阵. 输入是一个大小为$N\times D$的点云，其中，$N$为节点数，$D$是每个节点的特征维数. 首先，点云被输入到一个共享的MLP层，以提取每一个点的特征. 然后使用4个编码和解码层来学习每个点的特征. 最后，利用两层全连接层和一个退出层来预测每个点的语义标签信息并将其作为RandLA-Net的输出.

3.3 位姿估计模块

最后一个位姿估计模块用来评估图像像素局部特征与三维模型节点特征的相似性，以确定二维-三维对应. 将对应关系作为PnP方法的输入，使用RANSAC进行迭代优化，得到对象的最终位姿. 本文通过计算二维图像的像素级局部特征与三维网格模型的节点特征的相似性来建立二维-三维对应的匹配模型，给定像素级的局部特征矩阵$\boldsymbol{H}_m$和节点特征矩阵$\boldsymbol{H}_n$，

$$\boldsymbol{H}_m=\Lambda_m\left(\Phi_{\theta_1}(J)\right)\in\mathbf{R}^{|\mathcal{V}_m|\times D} \tag{6}$$

$$\boldsymbol{H}_n=\Upsilon_{\theta_2}\left(X_n,\mathcal{E}_n,\boldsymbol{U}_n\right)\in\mathbf{R}^{|\mathcal{V}_n|\times D} \tag{7}$$

由FCN中的Φ_{θ_1}和RandLA-Net中的Υ_{θ_2}分别通过式(8)计算互相关系得到特征相似性，即

$$\hat{S}=\mathrm{Softmax}\left(\boldsymbol{H}_m\boldsymbol{H}_n^{\mathrm{T}}\right)\in\mathbf{R}^{|\mathcal{V}_m|\times|\mathcal{V}_n|} \tag{8}$$

其中，Softmax函数是逐行执行，Λ_m收集预测为有效像素集合M_J所有相应的局部特征，$|\mathcal{V}_m|$表示有效像素数，$|\mathcal{V}_n|$是三维网格模型中的顶点数. 像素为有效值，表示它属于图像中的对象. 位姿估计网络的损失函数被定义为

$$\mathcal{L}(\theta)=\rho\mathrm{CEloss}\left(\mathrm{Softmax}\left(M_J\right),M_{\mathrm{gt}}\right)+\mathrm{CEloss}\left(\hat{S},S_{\mathrm{gt}}\right) \tag{9}$$

其中，CEloss表示交叉熵损失，$\theta=\{\theta_1,\theta_2\}$表示位姿估计网络的可训练参数，$\rho$是一个用来平衡这两部分损失的超参数. 损失的第一部分用于最小化模型的语义分割误差，而损失的第二部分则负责密集的二维-三维对应关系的匹配. 由于地面真实位姿在训练阶段是已知的，可以通过地面真实位姿将物体的三维模型投影到二维图像上来获得M_{gt}. S_{gt}表示使用3.4节中提出的算法获得的地面真实二维-三维对应关系.

3.4 二维-三维点向密集匹配

针对目标对象，本节使用的方法预测了语义分割掩码中每个像素的二维-三维对应. 这为位姿估计网络的训练提供了一个更稳定和鲁棒的约束，从而获得更准确的位姿估计结果. 本节将详细介绍获取二维-三维对应关系的方法. 像素级二维-三维密集匹配的主要原理如图6(b)所示，首先，找到包含该图像像素的所有三角形面片的边界虚线框. 其次，利用重心坐标[24]来区分三角形面片是否包含该图像像素，灰色的三角形面片在这个过程中被过滤掉. 最后，选择绿色三角形面片，其相对相机的深度最小，并将最近的顶点分配给该图像像素，将其作为一组二维-三维对应，如图7所示.

假设A，B和C表示投影到图像上的三角形面片的3个顶点，$p_i=\left(p_{ix},p_{iy},p_{iz}\right)$表示具有未确定深度$p_{iz}$的图像像素. 点$p_i$相对于三角形ABC的归一化重心坐标可以描述为$(1-a_2-a_3,a_2,a_3)$，系数$a_i$可表示为

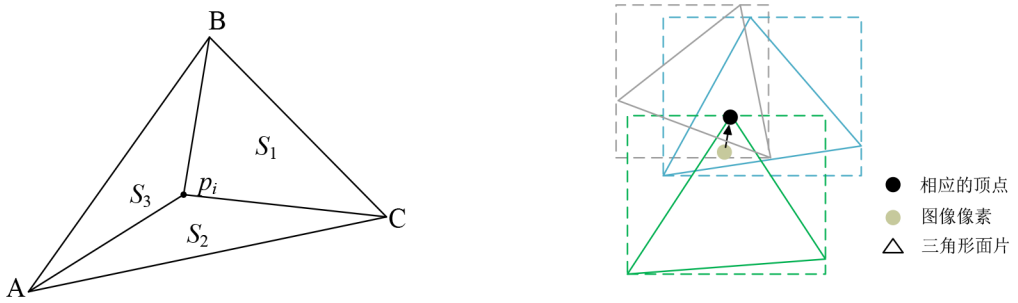
$$a_i=\frac{S_i}{S},i=1,2,3 \tag{10}$$

其中，S为给定三角形面积，S_i是第i个子三角形面积. 重心坐标原理如图6(a)所示.

如果像素p_i是三角块的重心，即p_i位于三角块ABC中，则重心坐标的所有元素都应该是非负的. 可以设

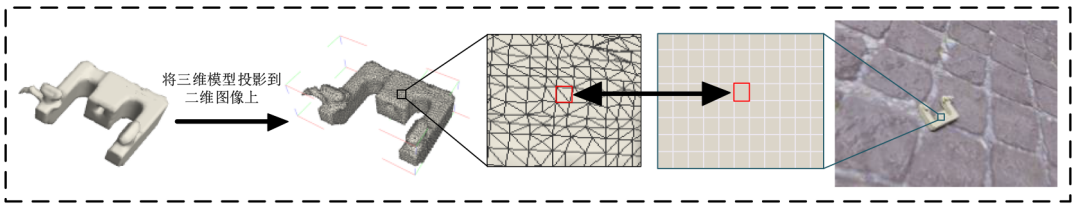
$$p_i=\left(1-a_2-a_3\right)A+a_2B+a_3C \tag{11}$$

或是以向量的形式，即



(a) 重心坐标原理　　(b) 像素分配原理

图6　二维平面中的重心坐标以及像素分配原理



(a) 重心坐标原理　　(b) 像素分配原理

图7　像素密集的二维-三维对应点匹配说明

$$[a_2 \quad a_3 \quad 1]\begin{bmatrix}\overrightarrow{AB}_x\\ \overrightarrow{AC}_x\\ \overrightarrow{p_iA}_x\end{bmatrix}=0,\ [a_2 \quad a_3 \quad 1]\begin{bmatrix}\overrightarrow{AB}_y\\ \overrightarrow{AC}_y\\ \overrightarrow{p_iA}_y\end{bmatrix}=0 \tag{12}$$

其中,$[a_2 \quad a_3 \quad 1]$分别与

$$\zeta_x=\left(\overrightarrow{AB}_x\overrightarrow{AC}_x\overrightarrow{p_iA}_x\right) \tag{13}$$

以及

$$\zeta_y=\left(\overrightarrow{AB}_y\overrightarrow{AC}_y\overrightarrow{p_iA}_y\right) \tag{14}$$

两两正交. 因此,可以得到

$$\mu[a_2 \quad a_3 \quad 1]=\zeta_x\times\zeta_y \tag{15}$$

其约束条件为

$$a_2\geqslant 0, a_3\geqslant 0, 1-a_2-a_3\geqslant 0 \tag{16}$$

其中,μ是线性关系标量. 对于每个边界框中包含p_i的面片,计算$\mathcal{T}_x$和$\mathcal{T}_y$的交叉积. 若结果满足式(15)的约束条件,则表示像素位于三角形面片中. 随后,根据式(16)进一步计算该补丁上的p_i深度,即

$$p_{iz}=\left(1-a_2-a_3\right)A_z+a_2B_z+a_3C_z \tag{17}$$

其中,A_z,B_z和C_z分别表示三角形面片的顶点的深度. 最后,选择具有最小p_{iz}的面片并将该面片中最近的顶点分配给p_i作为三维对应关系,整个过程如算法1所示.

算法1　为每个像素指定顶点

输入: 像素$p_i\in\mathcal{V}_m$以及网格模型的二维平面投影$T_{2d}=\{\mathcal{V}_n,\zeta_{\text{tri}}\}$,其中,$\zeta_{\text{tri}}$表示网格模型中的1个三角形面片

输出: 对于每个像素$C_{\text{pairs}}=\{(p_i,j)|p_i\in\mathcal{V}_m,j\in\mathcal{V}_n\}$的二维-三维对应

FOR each p_i in $\mathcal{V}_m$:

选择$\zeta_{\text{tri},i}=\{t_{ik}|t_{ik}\in\zeta_{\text{tri}},k=1,\cdots,l,\cdots\}$和$p_i$在$t_{ik}$的边界盒内;

选择$\tilde{\zeta}_{\text{tri},i}=\{t_{iq}|t_{iq}\in\zeta_{\text{tri},i},q=1,\cdots,m,\cdots\}$和$p_i$在$t_{iq}$的边界盒内带入式(15)和式(16)计算;

在$\tilde{\zeta}_{\text{tri},i}$中使用式(17)计算$p_i$的深度;

找出分配给p_i的最小深度的面片t_{is};

指定三角形面片t_{is}中最接近p_i的顶点j;

将p_i与j作为一组二维-三维对应点输出;

END FOR

3.5　训练超参数

本文利用PyTorch框架搭建网络结构,在训练和测试过程中,以单个RGB图像和点云模型作为输入,网络的输出为目标对象的3个旋转向量和3个平移向量,用于得到目标物体的旋转矩阵和平移矩阵. 在训练过程中损失函数中的超参数ρ设置为100,以平衡损失函数中两部分的大小. 该网络结构在NVIDIA RTX2080Ti GPU的Adam优化器上进行180个epoch的训练,每个小批量输入图像为16张图片,将前20个epoch的学习率设置为0.001,之后每20个epoch学习率减半,针对目标零件每10个epoch保存一次训练结果的权重模型.

4　实验结果与分析

为评估本文所提方法的有效性,利用工业零件模

型、LINEMOD数据集[16]和T-Less[25]里的部分模型生成的仿真数据集对位姿估计网络进行训练和测试. 仿真数据集是单个RGB目标图像数据集,每个RGB图像都有目标零件对象的旋转矩阵、平移矩阵和真实标签数据.

本节利用通用的几个度量标准来评估网络训练的效果,即5°5 cm度量、点的平均二维投影度量(2D Project metric)[26]和模型点的平均3D距离度量(Average 3D Distance,ADD)[27]等. 其中,点的平均二维投影度量是指三维网格模型在二维图像中的投影与地面真实姿态投影之间的平均距离. 从文献[6,8,12]可知,如果该距离小于5个像素,则认为该位姿是准确的. 模型点的平均3D距离度量是指利用网络估计出的姿态和地面真实姿态计算出的两个变换模型点之间的平均距离. 如果估计的位姿与变换后模型点云的真实位姿之间的平均距离小于对象直径的10%,则认为所估计的位姿正确. ADD指标的计算方式如下:

$$s=\frac{1}{|M|}\sum_{x_1\in M}\min_{M}\left(R_x+t\right)-\left(\hat{R}_x+\hat{t}\right) \tag{18}$$

其中,R和t是真实标签数据,$\hat{R}$和$\hat{t}$是网络预测的旋转和平移矩阵,M是三维模型的顶点坐标集.

4.1 仿真数据集的ADD测试结果

本节在仿真数据集上测试了伪孪生神经网络结构的有效性. 数据集分为训练和测试两部分,部分RGB图像用于训练网络,剩余图像用于测试. 表1展示了在ADD指标下的不同方法测试结果,使用本文方法所估计的几个目标对象中,花洒的位姿估计准确率最高为96.4%,其次是工业零件和相机分别为89.1%和88.6%,T-Less零件在本文方法下的ADD准确率达到90.2%. 低纹理工业零件由于结构相较于前几种更为复杂,其位姿估计的准确率低于前文所述部分目标对象. 对于同一个目标对象采用不同位姿估计网络训练,测试结果证明本文所运用的位姿估计精度优于部分主流方法.

图8给出了使用本文方法后部分仿真数据的目标物体位姿估计的可视化结果. 首先,从图中不难观察到位姿估计效果最好的依次是花洒、T-Less零件、工业零件、相机以及黄鸭,与表1中ADD测试结果保持一致. 其次,物体距离虚拟相机的远近并没有对位姿估计结果产生较大的影响,这也证明了本文网络中点向密集匹配方法的有效性.

表1 不同位姿估计方法在ADD指标下的测试结果 单位:%

估计对象	方法						
	PVNET	CDPN	YOLO6D	DPVL	DPOD	PVN3D	本文方法
工业零件	92.0	88.3	57.4	86.7	74.3	88.4	89.1
花洒	95.5	95.9	68.8	98.5	52.6	99.5	96.3
相机	86.9	91.7	36.6	94.1	24.2	99.6	88.6
黄鸭	52.6	66.8	27.2	63.5	26.1	98.2	67.2

4.2 消融实验

本节进行了消融实验[28],通过查阅文献[8~13,16,17]和工业零件数据集的训练与测试实验,对不同方法在不同度量指标下的性能表现进行了对照分析,具体内容如下表2所示. 第1行是5°5 cm测试的结果,除了文献[8]的方法之外,其他几种方法的准确率都保持了较高水平. 本文基于仿真数据集的训练将该指标提高至94.9%. 在二维投影度量中,由于工业零件结构的复杂性等因素的影响,虽然准确率不如部分对照方法但也达到了92.7%. 由于本文通过替换点云特征提取模块,降低了网络模型参数规模,在模型参数量指标上相较于前几种对照方法具有一定优势,此优势也进一步体现在了计算成本度量指标之中.

文献[8,10,11]中的方法都是通过回归每个像素的二维向量,与其类似的深度学习方法都可以作为某些特定特征的回归模型,而本文的方法训练一个深度神经网络来直接预测二维-三维对应. 如图9所示,以点的平均二维投影为度量标准(2D Project metric),从0像素到2像素为标准的几个位姿估计对象的准确率有明显的提升,虽然在5像素约束下,大部分的方法都能获得接近100%的二维投影度量,但当约束更为严格时,本文所提出的位姿估计方法的表现依然具有较强的鲁棒性.

4.3 与传统方法的对比实验

在真实场景下分割得到的单个点云模型,其点的个数往往是不确定的,而本节训练采用的数据集中目标对象是3 108个点. 因此,将多种采样点数的点云模型分别输入到训练好的权重模型中进行位姿预测,统计了各种方法在不同采样点数下的位姿估计准确率,

表2 工业零件数据集在不同度量指标以及方法下的测试对比结果 单位:%

度量指标	方法						
	PVNET	CDPN	YOLO6D	DPVL	DPOD	PVN3D	本文方法
5°5 cm	73.4	94.3	—	—	—	—	94.9
二维投影度量	99.0	98.1	90.4	99.4	—	—	92.7
网络参数量	12.9	26.7	7.8	—	13.5	13.2	10.1
计算成本	74.3	136.4	32.1	—	75.2	75.8	62.5

图8　目标对象位姿估计的可视化结果

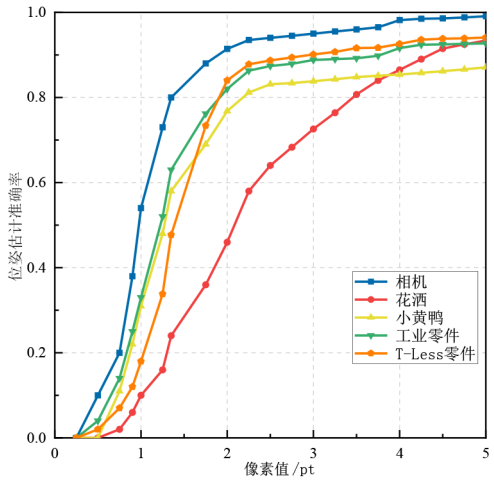


图9　不同目标对象下位姿估计的准确率

同时将本文方法与PVNet[8]方法及传统的点对特征匹配(PPF)方法进行对比.

图10是针对低纹理工业零件不同采样点数下位姿估计准确率的统计数据. 在不同的采样点数下,本文所

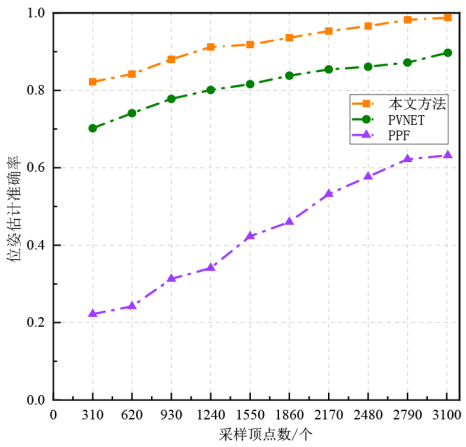


图10　不同采样点数下位姿估计的准确率

运用到的方法在估计位姿准确率上虽然有变化,但变化并不明显,说明所训练的模型可以针对不同点数的点云进行位姿估计并具有更高的精度. 同时,本文方法对每次取样点的位姿估计准确率都高于PVNet方法,对于传统的PPF方法,在采集点数变少时,其准确率出现了明显下降. 由此可见,使用本文及PVNet等深度学习方法来估计目标物体位姿的能力远优于一般的传统方法.

5 结论

本文运用一种适用于低纹理物体位姿估计的伪孪生神经网络,对于RGB和点云模型输入,通过全卷积网络(FCN)和三维点云语义分割网络(RandLA-Net)分别提取二维图像和三维模型的高维深层特征,使用网络推断密集的二维-三维对应关系并通过PNP方法求解出物体的6D位姿. 首先,采用物理仿真技术建立了大规模仿真数据集,解决了深度学习中大量数据集获取与标注较为繁琐的问题,对于每个估计对象,该数据集包含8 000张图片. 其次,通过仿真数据集训练6D位姿估计网络,证明了数据集的有效性,基于测试权重模型,模型点的平均3D距离度量准确率达到了89.1%,点的平均二维投影度量准确率达到了92.7%,验证了所用位姿估计网络方法的准确性和鲁棒性.

然而,本文方法局限于单一目标对象的图像处理,且需要训练特定的网络来进行位姿估计. 在后续研究工作中,将使用不同的方法扩展数据集,并提升网络在处理更加复杂场景图像时的能力.

参考文献

[1] XIANG Y, MOTTAGHI R, SAVARESE S. Beyond PASCAL: A benchmark for 3D object detection in the wild[C]// IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE, 2014: 75-82.

[2] HE Z X, WU C R, ZHANG S Y, et al. Moment-based 2.5-D visual servoing for textureless planar part grasping[J]. IEEE Transactions on Industrial Electronics, 2019, 66(10): 7821-7830.

[3] BORGHI G, VENTURELLI M, VEZZANI R, et al. POSEidon: face-from-depth for driver pose estimation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 5494-5503.

[4] MENG Y, LU Y, RAJ A, al et, Signet: Semantic instance aided unsupervised 3d geometry perception[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 9802-9812.

[5] XIANG Y, SCHMIDT T, NARAYANAN V, et al. PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes[EB/OL]. (2017-11-07)[2022-12]. https://arxiv.org/abs/1711.00199.

[6] KENDALL A, GRIMES M, CIPOLLA R. PoseNet: A convolutional network for real-time 6-DOF camera relocalization[C]//2015 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2015: 2938-2946.

[7] KEHL W, MANHARDT F, TOMBARI F, et al. SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again[C]//2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 1530-1538.

[8] PENG S D, LIU Y, HUANG Q X, et al. PVNet: Pixel-wise voting network for 6DoF pose estimation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway: IEEE, 2019: 4556-4565.

[9] YU X, ZHUANG Z Y, KONIUSZ P, et al. 6DoF object pose estimation via differentiable proxy voting loss[EB/OL]. (2020-02-10)[2021-12]. https://arxiv.org/abs/2002.03923.

[10] LI Z G, WANG G, JI X Y. CDPN: Coordinates-based disentangled pose network for real-time RGB-based 6-DoF object pose estimation[C]//2019 IEEE/CVF International Conference on Computer Vision(ICCV). Piscataway: IEEE, 2019: 7677-7686.

[11] ZAKHAROV S, SHUGUROV I, ILIC S. DPOD: 6D pose object detector and refiner[C]//2019 IEEE/CVF International Conference on Computer Vision(ICCV). Piscataway: IEEE, 2019: 1941-1950.

[12] WU C R, CHEN L, HE Z X, et al. Pseudo-Siamese graph matching network for textureless objects' 6-D pose estimation[J]. IEEE Transactions on Industrial Electronics, 2022, 69(3): 2718-2727.

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

[14] QI C R, YI L, SU H, et al. PointNet++: Deep hierarchical feature learning on point sets in a metric space[EB/OL]. (2017-06-07)[2021-12]. https://arxiv.org/abs/1706.02413.

[15] GAO G, LAURI M, HU X L, et al. CloudAAE: Learning 6D object pose regression with on-line data synthesis on point clouds[C]//2021 IEEE International Conference on Robotics and Automation. Piscataway: IEEE, 2021: 11081-11087.

[16] HE Y S, SUN W, HUANG H B, et al. PVN3D: A deep point-wise 3D keypoints voting network for 6DoF pose estimation[C]//2020 IEEE/CVF Conference on Computer

Vision and Pattern Recognition(CVPR). Piscataway: IEEE, 2020: 11629-11638.
[17] CHEN W, DUAN J M, BASEVI H, et al. PointPoseNet: Point pose network for robust 6D object pose estimation [C]//2020 IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE, 2020: 2813-2822.
[18] HE Y S, HUANG H B, FAN H Q, et al. FFB6D: A full flow bidirectional fusion network for 6D pose estimation [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway: IEEE, 2021: 3002-3012.
[19] KE Y, SUKTHANKAR R. PCA-SIFT: A more distinctive representation for local image descriptors[C]//Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2004, 2: II-II.
[20] LI S Q, XU C, XIE M. A robust O(n) solution to the perspective-n-point problem[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(7): 1444-1450.
[21] HU Q Y, YANG B, XIE L H, et al. RandLA-net: Efficient semantic segmentation of large-scale point clouds [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway: IEEE, 2020: 11105-11114.
[22] DENNINGER M, SUNDERMEYER M, WINKELBAUER D, et al. BlenderProc[EB/OL]. (2019-10-25) [2021-12]. https://arxiv.org/abs/1911.01911.
[23] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional Siamese networks for object tracking[EB/OL]. (2016-06-30)[2021-12]. https://arxiv.org/abs/1606.09549.
[24] SKALA V. Barycentric coordinates computation in homogeneous coordinates[J]. Computers & Graphics, 2008, 32(1): 120-127.
[25] HODAN T, HALUZA P, OBDRŽÁLEK Š, et al. T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects[C]//2017 IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE, 2017: 880-888.
[26] FEY M, LENSSEN J E, WEICHERT F, et al. SplineCNN: Fast geometric deep learning with continuous B-spline kernels[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 869-877.
[27] 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 the 11th Asian conference on Computer Vision. New York: ACM, 2012: 548-562.
[28] 左国玉, 张成威, 刘洪星, 等. 低质量渲染图像的目标物体6D姿态估计[J]. 控制与决策, 2022, 37(1): 135-141.
ZUO G Y, ZHANG C W, LIU H X, et al. 6D object pose estimation for low-quality rendering images[J]. Control and Decision, 2022, 37(1): 135-141.

作者简介

王神龙　男,1989年8月出生于安徽省安庆市. 上海理工大学机械工程学院副教授、硕士生导师. 主要研究方向为随机动力学与控制、机器视觉与机器学习等.
E-mail: shenlongwang@usst.edu.cn

雍　宇　男,1996年11月出生于江苏省扬州市.上海理工大学机械工程学院硕士研究生. 研究方向为目标检测、6D位姿估计.
E-mail: 1219817191@qq.com

吴晨睿(通讯作者)　男,1989年9月出生于黑龙江省哈尔滨市.上海理工大学机械工程学院讲师、硕士生导师.主要研究方向为机器人视觉伺服控制、工业零件目标位姿估计等.
E-mail: wuchenrui@usst.edu.cn