引用格式：王立春，杨超，付芳玉. 结构化约束增强的6D物体位姿估计[J]. 北京工业大学学报，2025，51(2)：173-182.
WANG L C，YANG C，FU F Y. 6D object pose estimation enhanced by structural constraint[J]. Journal of Beijing University of Technology，2025，51(2)：173-182.（in Chinese）

结构化约束增强的6D物体位姿估计

王立春[1,2]，杨 超[1,2]，付芳玉[1,2]
（1. 北京工业大学信息学部，北京 100124；2. 北京工业大学多媒体与智能软件技术北京市重点实验室，北京 100124）

摘 要：针对基于投票策略的6D物体位姿估计方法忽略了关键点间结构信息的问题，提出结构化约束增强的6D物体位姿估计方法——SC-Pose。该方法定义了一种用于描述物体2D关键点间结构化信息的形状描述符，通过增加关键点结构化损失约束形状描述符的预测值与真值相近，从而使2D关键点的定位更加准确，提升了6D物体位姿估计的精度。在LINEMOD、OCC-LINEMOD和TruncationLINEMOD数据集上进行了实验，结果表明，SC-Pose方法可以明显提升6D物体位姿估计的性能。

关键词：6D物体位姿估计；单位向量场；投票策略；结构化损失；抓取交互；深度网络
中图分类号：TP 391 **文献标志码**：A **文章编号**：0254－0037(2025)02－0173－10
doi：10.11936/bjutxb2023040019

6D Object Pose Estimation Enhanced by Structural Constraint

WANG Lichun[1,2]，YANG Chao[1,2]，FU Fangyu[1,2]
（1. Faculty of Information Technology，Beijing University of Technology，Beijing 100124，China；
2. Beijing Key Laboratory of Multimedia and Intelligent Software Technology，
Beijing University of Technology，Beijing 100124，China）

Abstract：Aiming at the problem that the 6D object pose estimation method based on the voting strategy ignores the structural information between keypoints，a 6D object pose estimation method enhanced by structural constraint（SC-Pose）is proposed. This method defines a shape descriptor designed to describe the structured information between the 2D keypoints of the object. By enhancing the keypoint structural loss to constrain the predicted shape descriptor to be close to the ground-truth shape descriptor，the positioning of the 2D keypoints is more accurate，and thereby ultimately e incorporating the accuracy of 6D object pose estimation. Results on the LINEMOD，OCC-LINEMOD and TruncationLINEMOD datasets show that SC-Pose can significantly boost the accuracy of 6D object pose estimation.

Key words：6D object pose estimation；unit vector-field；voting strategy；structural loss；grasping interaction；deep network

从相机获取的RGB图像中估计6D物体位姿在自动驾驶、工业机器人以及增强现实等领域具有广泛应用，但是在现实场景中，受光照变化、物体遮挡及纹理信息不足、场景杂乱等因素的影响，准确地从

收稿日期：2023-04-11；修回日期：2023-05-22
基金项目：国家自然科学基金资助项目(62376014)；中国高校产学研创新基金资助项目(2021JQR023)
作者简介：王立春(1975—)，女，教授，主要从事人工智能、场景理解方面的研究，E-mail：wanglc@bjut.edu.cn

RGB 图像中估计 6D 物体位姿成为一个具有挑战的问题。基于对应关系的 6D 物体位姿估计方法(BB8[1]、Yolo6D[2]等)先建立 2D 图像关键点与 3D 模型关键点之间的对应关系,再通过几何方法求解 6D 物体位姿参数,通过这种方式计算物体位姿仅需要少量的关键点的对应关系,就可以快速、准确地估计物体位姿。

传统的基于对应关系的方法从 RGB 图像中手工提取特征用于构建 2D 关键点与 3D 关键点之间的对应关系,由于依赖手工提取的特征,这类方法对于复杂场景不鲁棒。受益于深度学习的发展,通过基于投票策略的方法建立关键点的对应关系取得了比传统方法更好的位姿估计性能。

基于投票策略的 6D 物体位姿估计方法使深度网络更多地关注目标物体的局部特征,从而减轻了杂乱背景的影响,实现了较高的物体位姿估计性能。其一般流程为:首先,通过深度网络为 3D 关键点在图像上的投影点,即 2D 关键点 m_j,预测单位向量场,单位向量场中的单位向量表示目标物体像素点指向 2D 关键点的方向;然后,基于单位向量场任意选取 2 个单位向量来计算其交点,并将交点作为假设关键点,重复 N 次,获得假设关键点集合;最后,通过目标物体所有像素点对假设关键点投票得到每个假设关键点的置信度分数,将置信度分数最高的假设关键点作为预测的 2D 关键点 m_j',从而建立 3D 关键点集合$\{M_j\}$与预测的 2D 关键点集合$\{m_j'\}$中关键点的一一对应关系,再利用透视 n 点(perspective-n-point, PnP)[3]算法计算物体 6D 位姿。

在以上计算 6D 物体位姿的过程中,2D-3D 关键点的对应关系是 6D 物体位姿估计重要的中间表示方式,2D 关键点的准确性直接影响 6D 物体位姿估计的准确性。目前的方法需要预测的 2D 关键点 m_j' 尽可能接近 m_j,这就要求预测的 2D 关键点有尽可能小的位置误差。由于 2D 关键点是由标定的 3D 关键点投影得到的,而标定的 3D 关键点是 3D 模型上的点,因此,关键点之间的拓扑结构可以在一定程度上描述物体的形状。约束预测关键点的拓扑结构与真实的拓扑结构相似,可以约束多个预测关键点的位置同时满足一定的精度要求,从而影响位姿估计的准确性。因此,本文提出结构化约束增强的 6D 物体位姿估计方法——SC-Pose。该方法借助物体 2D 关键点间的几何结构辅助 2D 关键点进行准确定位,最终提升了 6D 物体位姿估计性能。本文在 3 个 6D 位姿估计任务公共数据集上进行广泛实验,实验结果证明了本文方法的有效性。最后,基于本文方法模型进行机器人抓取交互实验,实验结果证明了 SC-Pose 在真实场景中具有良好的应用效果,可以部署在机器人抓取任务中。

1 相关工作

6D 物体位姿估计的目的是估计物体坐标系到相机坐标系的旋转变换和平移变换,即估计 3D 旋转矩阵 $\boldsymbol{R}$ 和 3D 平移向量 $\boldsymbol{t}$。现有的基于 RGB 图像的 6D 物体位姿估计方法可分为基于对应关系的方法、基于模板的方法和基于投票策略的方法。

1) 基于对应关系的方法。基于对应关系的方法需要预先对物体 3D 模型(见图 1)标注 K 个 3D 关键点$\{M_j \mid j=1,2,\cdots,K\}$以及 3D 模型质心,然后,从不同的视角采集模型图像,得到 3D 关键点在图像上的投影点,即 2D 关键点,从而建立关键点的 3D 模型坐标与 2D 像素坐标的一一对应关系。最后,将这些对应的 3D 关键点与 2D 关键点输入 PnP 算法中计算物体 6D 位姿。

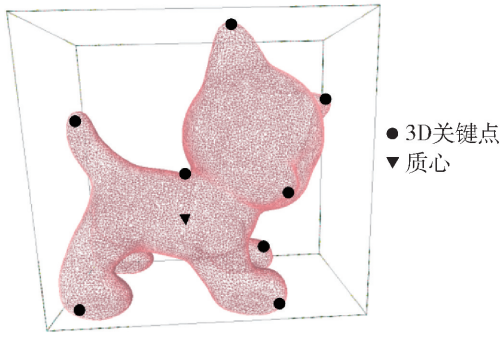


图 1 物体 3D 模型

Fig. 1 3D object model

以 RGB 图像为输入的研究工作中,一些研究工作仅建立稀疏的 2D-3D 关键点的对应关系,这些研究工作将物体 3D 模型包围框的 8 个顶点定为 3D 关键点,通过深度网络预测 3D 关键点在图像上的投影点,从而建立 2D-3D 关键点的对应关系,代表方法有 BB8[1]、Yolo6D[2]等。然而,有研究工作认为将物体 3D 模型包围框的 8 个顶点作为 3D 关键点是不合理的,因为 3D 模型包围框顶点在图像上的投影点远离图像中目标物体像素,距离物体像素越远,包围框顶点在图像上的投影点的定位误差越大。因此,一些研究工作提出在物体 3D 模型表面定义 3D 关键点,代表方法有 PVNet[4]、HybridPose[5]、RPVNet[6]、KDFNet[7]和 DPVR[8]等。还有一些研究工作预测密集的 3D 关键点的投影

点，而不是稀疏的投影点，从而建立密集的 2D-3D 关键点的对应关系，代表方法有 DPOD[9]、Pix2Pose[10]、KeyPose[11] 和 EPOS[12] 等。总体而言，基于对应关系的方法需要物体具有丰富纹理信息才可以获得准确的 2D-3D 关键点的对应关系，对于弱纹理或者无纹理物体，此类方法的位姿估计效果较差。

2）基于模板的方法。基于模板的方法常被应用于处理弱纹理或者无纹理物体。该方法是指已知物体 3D 模型在不同角度的观测视图以及视图中物体真实位姿，然后匹配当前预测视图与已知的观测视图，将观测视图中的 6D 物体位姿作为当前视图的物体位姿。简而言之，该类方法首先要建立模板库，从标记有真实 6D 位姿的模板库中找到最相似的视图模板，将该视图模板的真实位姿作为当前视图中目标物体的位姿。

以 RGB 图像为输入的研究工作中，一类方法为直接从 RGB 图像中预测目标物体的 6D 位姿，这个过程可以看作从已训练的带有真实位姿的 RGB 图像中，寻找与当前输入图像最近似的图像，并且输出其对应的 6D 物体位姿作为位姿估计预测结果，代表方法有 PoseCNN[13]、SSD-6D[14]、Deep-6DPose[15]、基于坐标的分离位姿网络（coordinates-based disentangled pose network，CDPN）[16]、Robust-6DPose[17] 等；另一类方法为针对一类物体构建隐式表示，然后通过深度网络输出预测位姿，代表方法有归一化对象坐标空间（normalized object coordinate space，NOCS）[18]、SPD-6DPose[19]、Latentfusion[20]、LCSS-6DPose[21] 等。总体而言，基于模板的方法可以处理弱纹理以及无纹理物体，但该类方法也存在问题，即：在位姿估计之前需要建立模板库，较为费时费力，并且基于模板的位姿估计方法较难估计场景中被遮挡物体的位姿。

3）基于投票策略的方法。基于投票策略的方法常被应用于处理遮挡场景的 6D 物体位姿估计，因为当物体一部分被遮挡时，基于模板的方法可能遇到困难。该方法是指图像上目标物体的每个像素点通过投票对最终预测的 6D 位姿产生影响，该方法可以有效地应对遮挡情况。

基于投票策略的方法首先通过深度网络对输入图像进行分割以获得目标物体区域，然后使用投票方式预测图像上目标物体的 2D 关键点，从而建立稀疏的 2D-3D 关键点的对应关系，最后使用 P*n*P 算法计算物体 6D 位姿，代表方法有 PVNet[4]、DPVR[8] 等。还有一些基于投票策略的方法通过预测密集的 2D-3D 关键点的对应关系提高 6D 物体位姿估计准确性，代表方法有 T6D-Direct[22]、EfficientPose[23] 等。

2　方法

SC-Pose 的整体框架如图 2 所示。首先，SC-Pose 利用深度网络对输入的 RGB 图像进行语义分割和单位向量场预测，基于语义分割结果以及预测的单位向量场任意选取 2 个单位向量计算其交点，并将交点作为假设关键点，重复 N 次，获得假设关键点集合；其次，目标物体区域内的所有像素点为假设关键点投票以获得预测的 2D 关键点；然后，利用 SC-Pose 提出的方法建立 2D 关键点间的几何结构，并利用提出的基于形状描述符的结构化损失进行训练。由于物体模型上的 3D 关键点已知，从而可以得到预测的 2D 关键点和 3D 关键点的一一对应关系，最后使用 P*n*P 算法计算物体 6D 位姿。

2.1　基于 2D 关键点的形状描述符

本文借助骨架的思想[24]，基于 2D 关键点以及 3D 模型质心在图像上的投影点 o 定义一种用于描述图像中物体形状的描述符。为方便起见，3D 模型质心在图像上的投影点 o 称为物体中心点。形状描述符真值和预测值如图 3 所示。

对有真值 2D 关键点 $\{m_j\}$ 的图像 I，连接物体中心点 o 与关键点 m_j 得到一组线段 $\{om_j \mid 1 \leqslant j \leqslant K\}$，再将所有关键点首尾相连得到另一组线段 $\{m_jm_{j+1} \mid 1 \leqslant j < K\} \cup \{m_Km_1\}$，则这 2 组线段的并集构成图像中目标物体的形状描述符 $S_{sd} = \{om_j \mid 1 \leqslant j \leqslant K\} \cup \{m_jm_{j+1} \mid 1 \leqslant j < K\} \cup \{m_Km_1\}$，如图 3（a）所示。同样的方式，基于网络预测的 2D 关键点 $\{m'_j\}$ 以及物体中心点 o 可以得到对物体形状描述符的预测 $S'_{sd} = \{om'_j \mid 1 \leqslant j \leqslant K\} \cup \{m'_jm'_{j+1} \mid 1 \leqslant j < K\} \cup \{m'_Km'_1\}$，如图 3（b）所示。

2.2　网络结构

SC-Pose 需要进行像素级别的网络预测，这样可以使网络在处理被遮挡物体时更加关注图像的局部区域。为了提升特征提取的能力，可以使用深层次的网络来提取特征，但是随着网络深度的增加，梯度消失和梯度爆炸的情况随之出现，虽然存在一些特殊技巧（如正则化、随机失活）可以缓解梯度问题，但同时也会带来网络退化的问题。为了既能缓解梯度问题，又能解决网络退化问题，SC-Pose 使用 ResNet 残差网络[25] 作为主干网络来提升特征提取

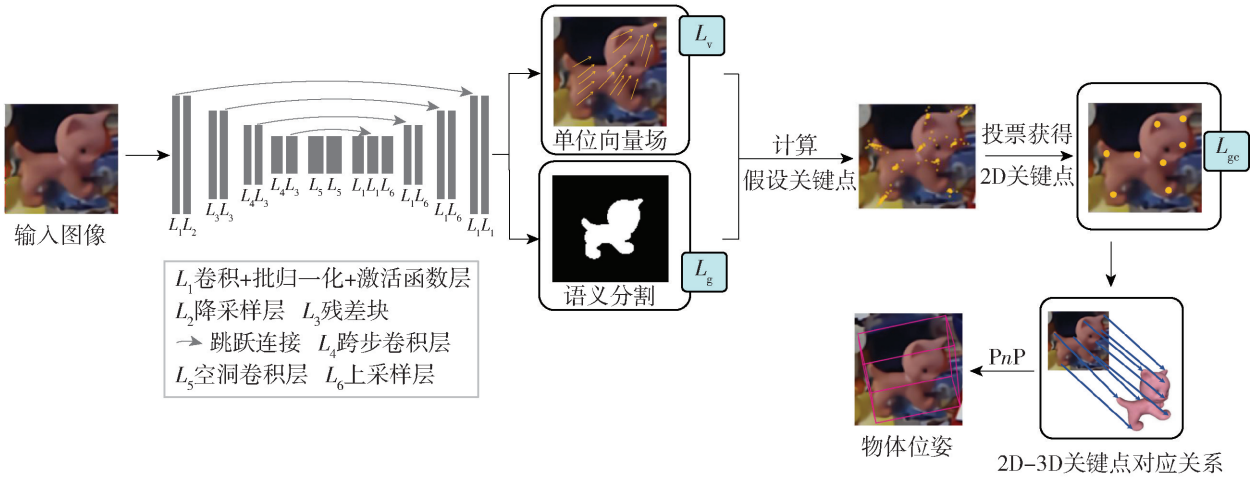
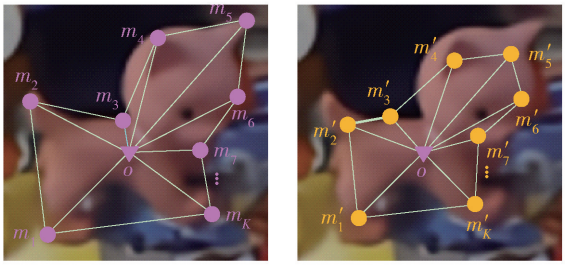


图 2 SC-Pose 框架

Fig. 2 SC-Pose framework



(a) 形状描述符真值 (b) 形状描述符预测值

图 3 形状描述符真值和预测值

Fig. 3 Ground-truth shape descriptor and predicted shape descriptor

能力。

SC-Pose 网络的输入为单张 RGB 图像 $I \in R^{H\times W\times 3}$,网络输出为语义分割结果 $S \in R^{H\times W\times 2}$ 和针对 2D 关键点 m_j 预测的单位向量场 $\{\boldsymbol{V}(m_j) \in R^{H\times W\times 2} \mid j=1,2,\cdots,K\}$,$R$ 表示特征空间。SC-Pose 网络的组成成分有卷积层、批归一化层、激活函数层、降采样层、跨步卷积层、上采样层以及不同层次的跳跃连接残差块。首先,网络对输入的 $H\times W\times 3$ 尺寸的 RGB 图像执行卷积和池化操作;然后,在特征图上执行跳跃连接和上采样,直到特征图尺寸达到 $H\times W$;最后,对 $H\times W$ 的特征图应用 1×1 卷积核提升维度输出预测的单位向量场 $\{\boldsymbol{V}(m_j)\}$ 以及预测的语义分割结果 S。基于单位向量场 $\boldsymbol{V}(m_j)$ 可以得到图像中像素点 p 指向 2D 关键点 m_j 的单位向量 $\boldsymbol{v}_j(p)$,基于语义分割结果 S 可以得到目标物体像素点集合 E。

在计算假设关键点时,为每个假设关键点计算置信度分数,计算方式为

$$\omega_j^i(h_j^i)=\sum_{p\in E} I\left(\frac{h_j^i-p}{\|h_j^i-p\|_2}\cdot \boldsymbol{v}_j(p)\geqslant\varphi\right) \tag{1}$$

式中:p 为目标物体像素点集合 E 中的像素点;h_j^i 为 2D 关键点 m_j 的第 i 个假设关键点;I 为指示函数;φ 为阈值,在实验中设置为 0.99;$\|\cdot\|_2$ 表示欧几里得范数,本文中用于计算像素 p 指向假设关键点 h_j^i 的向量长度。

假设关键点的置信度分数越高,意味着有越多的预测单位向量指向该假设关键点。选取置信度分数最高的假设关键点作为预测的 2D 关键点 m_j',即 $m_j'=\underset{h_j^i}{\operatorname{argmax}}(\omega_j^i(h_j^i))$,$1\leqslant j\leqslant K$,$0\leqslant i\leqslant N-1$。

2.3 基于形状描述符的结构化损失

在以往的方法中,网络约束物体上每个关键点的位置尽量准确,忽略了关键点间相对位置对预测关键点精度的影响。本文利用形状描述符度量关键点间的相对位置,定义结构化损失为

$$L_{\mathrm{gc}}=\sum_{j=1}^{K}\left(1-\frac{\boldsymbol{w}_j'\cdot\boldsymbol{w}_j}{|\boldsymbol{w}_j'|\times|\boldsymbol{w}_j|}\right) \tag{2}$$

式中:K 为 2D 关键点数量;$\boldsymbol{w}_j$ 为基于 S_{sd} 定义的特征向量,$\boldsymbol{w}_j=[d_j,g_j,g_{j+1}]$,$d_j$ 为线段 m_jm_{j+1} 的长度,$j=K$ 时,d_j 为 m_Km_1 的长度,g_j 为线段 om_j 的长度,$j=K$ 时,$\boldsymbol{w}_K=[d_K,g_K,g_1]$;$\boldsymbol{w}_j'$ 为基于 S_{sd}' 定义的特征向量,$\boldsymbol{w}_j'=[d_j',g_j',g_{j+1}']$,$d_j'$ 为线段 $m_j'm_{j+1}'$ 的长度,$j=K$ 时,d_j' 为线段 $m_K'm_1'$ 的长度,g_j' 为线段 om_j' 的长度,$j=K$ 时,$\boldsymbol{w}_K'=[d_K',g_K',g_1']$。通过约束使 $\boldsymbol{w}_j$ 和 $\boldsymbol{w}_j'$ 相近,借助其他 2D 关键点信息,使预测的 2D 关键点的位置偏差减小,进而提升位姿估计的性能。

2.4 网络总损失

网络总损失的定义为

$$L = L_{\mathrm{g}} + L_{\mathrm{v}} + \lambda L_{\mathrm{gc}} \tag{3}$$

式中：λ 为经验参数，本文实验中设置为 0.1；L_{g} 为语义分割损失；L_{v} 为单位向量场预测损失；L_{gc} 为本文提出的基于形状描述符的结构化损失。

语义分割损失 L_{g} 采用交叉熵损失，具体计算公式为

$$L_{\mathrm{g}} = -a_p \ln a'_p - (1-a_p)\ln(1-a'_p) \tag{4}$$

式中：a'_p 为像素点 p 被网络预测为目标物体像素点的概率；a_p 为像素点 p 的语义真值。

单位向量场预测损失 L_{v} 的计算公式为

$$L_{\mathrm{v}} = \sum_{j=1}^{K}\sum_{p\in E}\left(l_1(\boldsymbol{v}_j^{\mathrm{ho}}(p) - \boldsymbol{u}_j^{\mathrm{ho}}(p)) + l_1(\boldsymbol{v}_j^{\mathrm{ve}}(p) - \boldsymbol{u}_j^{\mathrm{ve}}(p))\right) \tag{5}$$

式中：p 为目标物体像素集合 E 中的像素；$\boldsymbol{v}_j^{\mathrm{ho}}$、$\boldsymbol{v}_j^{\mathrm{ve}}$ 分别为预测单位向量 $\boldsymbol{v}_j(p)$ 沿水平方向和竖直方向的分量，$\boldsymbol{v}_j^{\mathrm{ho}}(p) = \boldsymbol{V}(m_j)[p_w, p_h, 0]$，$\boldsymbol{v}_j^{\mathrm{ve}}(p) = \boldsymbol{V}(m_j)[p_w, p_h, 1]$；$\boldsymbol{u}_j^{\mathrm{ho}}(p)$ 和 $\boldsymbol{u}_j^{\mathrm{ve}}(p)$ 分别为真实单位向量 $\boldsymbol{u}_j(p)$ 沿水平和竖直方向的分量；$l_1(\,)$ 为 smooth l_1 函数。

3　实验

3.1　数据集

为了评估 6D 物体位姿估计方法的性能，研究人员提出了各种数据集，这些数据集包含了大量的 RGB 图像和物体的 6D 物体位姿信息，不仅覆盖了不同类型和数量的物体，还提供了不同环境影响的挑战，从而帮助测试 6D 物体位姿估计方法的性能。

本文在常用的 6D 物体位姿估计数据集 LINEMOD、OCC-LINEMOD 和 TruncationLINEMOD 上进行实验，并与其他优秀方法进行对比。

LINEMOD 数据集[26]是一个被广泛应用于物体 6D 物体位姿估计的经典数据集。该数据集包含 13 个弱纹理的物体，其中每个物体大约有 1 200 张图像，共 15 783 张图像，每张图像的分辨率为 640 × 480 像素。构建该数据集时场景较为杂乱，并且光照变化幅度较大，导致对该数据集图像上的物体进行 6D 位姿估计十分困难。本文遵循文献[4]的工作将该数据集中 15% 的样本用于训练，85% 的样本用于测试，其中每个物体大约有 180 张图像用于训练，超过 1 000 张图像用于测试。

OCC-LINEMOD 数据集[27]是 LINEMOD 数据集的子集，主要关注被遮挡的物体。该数据集中的每张图像包含多个物体，并且这些物体都受到严重遮挡。OCC-LINEMOD 数据集中的物体种类与 LINEMOD 数据集相同，但在遮挡情况下物体的形状、外观等信息难以获取，导致物体的 2D 关键点难以定位，从而使得估计该数据集中物体的位姿更加困难。因此，对于 6D 物体位姿估计算法的评估，OCC-LINEMOD 数据集提供了更加严格和真实的测试条件。

TruncationLINEMOD 数据集[4]是通过随机裁剪 LINEMOD 数据集中的图像创建的，裁剪后只有 40% ~60% 的目标物体区域可见，在该数据集上进行实验可以充分评估 6D 物体位姿估计方法的预测效果。

3.2　训练策略

在网络训练过程中，本文以单张 RGB 图像作为网络输入，图像的分辨率为 640 × 480 像素，Batchsize 设置为 16，初始学习率设置为 0.001。为了控制模型的训练速度，每间隔 20 个 epoch，学习率衰减为上一次的 1/2。此外，每间隔 3 个 epoch 对模型进行评估。

为了避免训练模型发生过拟合现象，本文遵循文献[4]的工作使用合成数据。具体来说，本文在 LINEMOD 数据集中增加渲染以及合成图像，对每类物体从不同的视点渲染 10 000 张图像。此外，通过背景替换方法，将 SUN397[28]数据集的图像作为物体背景，合成 10 000 张图像。这些合成数据可以增加模型的训练样本数量和多样性，并且能够模拟真实场景中的光照和物体背景，有助于提高模型的泛化能力。为了提高训练模型的鲁棒性，本文在训练过程中使用随机旋转、随机裁剪、添加噪声和颜色抖动等数据增强方法，这些方法可以使模型在面对未知场景时表现更好，从而提高模型的实用性。

3.3　评价指标

平均距离度量（average distance metric，ADD）是指对物体 3D 模型顶点 x 分别使用估计位姿（$\boldsymbol{R}$，$\boldsymbol{t}$）和真实位姿（$\hat{\boldsymbol{R}}$，$\hat{\boldsymbol{t}}$）进行变换，并计算变换后 2 个点集中对应点之间的平均距离。当平均距离小于物体模型直径（2 个点集中对应点间的最大距离称为模型直径）的 10% 时，估计的物体位姿被认为是正确的，对于非对称物体，使用 ADD 指标，具体计算方式为

$$M_{\mathrm{ADD}} = \frac{1}{q}\sum_{x\in Q}\left\|(\hat{\boldsymbol{R}}x + \hat{\boldsymbol{t}}) - (\boldsymbol{R}x + \boldsymbol{t})\right\|_2 \tag{6}$$

式中：Q 为 3D 模型点集合；q 为 3D 模型点数量，$\hat{\boldsymbol{R}}$、$\hat{\boldsymbol{t}}$ 为真实旋转矩阵和平移向量；$\boldsymbol{R}$、$\boldsymbol{t}$ 为 SC-Pose 估计的旋转矩阵和平移向量。

对于对称物体,由于位姿的歧义性,使用用于对称物体的平均距离度量(average distance metric for symmetric, ADD-S)指标,其中平均距离根据最近点计算,计算公式为

$$M_{\text{ADD-S}}=\frac{1}{q}\sum_{x_1\in Q,x_2\in Q}\min\|(\hat{\boldsymbol{R}}x_1+\hat{\boldsymbol{t}})-(\boldsymbol{R}x_2+\boldsymbol{t})\|_2 \tag{7}$$

3.4 位姿估计实验

本文在常用的 6D 位姿估计数据集上进行了广泛的实验,以验证本文提出的 SC-Pose 的有效性,同时,将 SC-Pose 与基于 RGB 图像输入的其他有竞争力的 6D 位姿估计方法进行对比。

1)在 LINEMOD 数据集上的实验结果分析

SC-Pose 在 LINEMOD 数据集上的 ADD(-S)指标上与其他方法的对比实验结果如表 1 所示。表中:“*”表示对称物体;ADD(-S)指对于非对称物体使用 ADD 指标,对于对称物体使用 ADD-S 指标;粗体数值表示最优结果。从表 1 中可以看出,与其他方法相比,SC-Pose 在 LINEMOD 数据集的大多数物体类别上取得了最优的位姿估计结果,平均准确率实现了至少 4.63% 的增幅,说明通过约束预测的关键点拓扑结构与真实拓扑结构相近,可以在预测 2D 关键点时利用其他 2D 关键点的信息减小自身的定位误差,从而提高位姿估计的准确率。

表 1 在 LINEMOD 数据集上的 ADD(-S)指标准确率对比

Table 1 Accuracy comparison in terms of ADD(-S) on LINEMOD dataset %

方法	Bb8[1]	Yolo6D[2]	PVNet[4]	SSD-6D[14]	Brachman[27]	SC-Pose
Ape	40.0	21.6	43.62	65	33.2	**67.14**
Bench	91.8	81.8	99.90	80	64.8	**100.00**
Cam	55.7	36.6	86.86	78	38.4	**89.22**
Can	64.1	68.8	95.47	86	62.9	**96.06**
Cat	62.6	41.8	79.34	70	42.7	**86.23**
Driller	74.4	63.5	96.43	73	61.9	**98.22**
Duck	44.3	27.2	52.58	**66**	30.2	63.76
Eggbox*	57.8	69.6	99.15	**100**	49.9	99.25
Glue*	41.2	80.0	95.66	**100**	31.2	98.75
Holep	67.2	42.6	81.92	49	52.8	**89.25**
Iron	84.7	74.9	98.88	78	80.0	**99.18**
Lamp	76.5	71.1	99.33	73	67.0	**99.90**
Phone	54.0	47.7	92.41	79	38.1	**94.72**
平均值	62.7	55.9	86.27	79	50.2	**90.90**

2)在 OCC-LINEMOD 数据集上的实验结果分析

SC-Pose 在 OCC-LINEMOD 数据集上的 ADD(-S)指标的对比实验结果如表 2 所示。

从表 2 可以看出,SC-Pose 在预测物体位姿较为困难的数据集上实现了最优的位姿估计性能,与对比方法相比,平均准确率实现了至少 4.01% 的性能提升。因为 OCC-LINEMOD 数据集中的物体遭受严重的遮挡,所以 SC-Pose 在该数据集上的性能提升幅度小于在 LINEMOD 数据集上的位姿估计的性能提升。

SC-Pose 在 TruncationLINEMOD 数据集上的 ADD(-S)指标的对比实验结果如表 3 所示。可以看出,SC-Pose 在大多数类别上都要优于对比方法,与其他方法相比,平均准确率实现了至少 3.89% 的增幅,达到了最高的平均准确率。该数据集中的图像只有 40% ~60% 的目标物体区域可见,因此,在该数据集上预测物体位姿十分困难,实验结果证明了 SC-Pose 的有效性。

表 2　在 OCC-LINEMOD 数据集上的 ADD(-S) 指标准确率对比

Table 2　Accuracy comparison in terms of ADD(-S) on OCC-LINEMOD dataset　%

模型	PVNet[4]	RPVNet[6]	Pix2Pose[10]	SC-Pose
Ape	15.8	17.9	**22.0**	21.11
Can	63.3	69.5	44.7	**74.57**
Cat	16.7	19.0	22.7	**23.34**
Driller	65.7	63.7	44.7	**73.15**
Duck	25.2	31.1	15.0	**37.16**
Eggbox*	50.2	**59.2**	25.2	50.21
Glue*	49.6	46.6	32.4	**55.81**
Holep	39.7	42.8	**49.5**	46.36
平均值	40.8	43.7	32.0	**47.71**

表 3　在 TruncationLINEMOD 数据集上的 ADD(-S) 指标准确率对比

Table 3　Accuracy comparison in terms of ADD(-S) on TruncationLINEMOD dataset　%

模型	PVNet[4]	RPVNet[6]	SC-Pose
Ape	12.8	14.1	**14.48**
Bench	42.8	41.7	**50.78**
Cam	27.7	27.3	**29.97**
Can	32.9	32.6	**42.14**
Cat	25.2	24.9	**27.41**
Driller	37.0	37.2	**42.75**
Duck	12.4	**14.8**	13.72
Eggbox*	44.1	43.5	**45.57**
Glue*	38.1	**38.4**	37.78
Holep	22.4	23.2	**28.13**
Iron	42.0	40.9	**45.40**
Lamp	40.9	38.8	**48.21**
Phone	30.9	30.0	**33.76**
平均值	31.5	31.3	**35.39**

3.5　消融实验

本文在位姿估计较为困难的 OCC-LINEMOD 数据集和 TruncationLINEMOD 数据集上进行消融实验，实验结果如表 4、5 所示。表中的 Baseline 方法不构建形状描述符，不使用基于形状描述符的结构化损失。可以看出，相比于 Baseline 方法，本文方法在 OCC-LINEMOD 数据集和 TruncationLINEMOD 数据集上分别实现了 4.77% 和 4.07% 的精度提升，证明了本文方法在位姿估计方面的有效性。

表 4　在 OCC-LINEMOD 数据集上的 ADD(-S) 指标准确率消融实验

Table 4　Ablation experiment on the accuracy of ADD(-S) on OCC-LINEMOD dataset　%

模型	Baseline	SC-Pose
Ape	18.12	**21.11**
Can	69.01	**74.57**
Cat	19.55	**23.34**
Driller	70.68	**73.15**
Duck	26.10	**37.16**
Eggbox*	41.49	**50.21**
Glue*	51.60	**55.81**
Holep	**46.95**	46.36
平均值	42.94	**47.71**

表 5　在 TruncationLINEMOD 数据集上的 ADD(-S) 指标准确率消融实验

Table 5　Ablation experiment on the accuracy of ADD(-S) on TruncationLINEMOD dataset　%

模型	Baseline	SC-Pose
Ape	12.86	**14.48**
Bench	42.80	**50.78**
Cam	27.89	**29.97**
Can	33.11	**42.14**
Cat	24.68	**27.41**
Driller	36.44	**42.75**
Duck	11.24	**13.72**
Eggbox*	44.05	**45.57**
Glue*	**38.11**	37.78
Holep	22.40	**28.13**
Iron	41.84	**45.40**
Lamp	40.91	**48.21**
Phone	30.86	**33.76**
平均值	31.32	**35.39**

3.6　可视化实验

为了进一步证明 SC-Pose 的有效性，本文可视化了 SC-Pose 预测的物体位姿，如图 4～6 所示。图中物体的 6D 位姿用 3D 包围框表示，3D 包围框的中心点表示物体的位置，3D 包围框的朝向表示物体的方向。

图 4 LINEMOD 数据集位姿可视化

Fig. 4 Visualization of poses on LINEMOD dataset

图 5 OCC-LINEMOD 数据集位姿可视化

Fig. 5 Visualization of poses on OCC-LINEMOD dataset

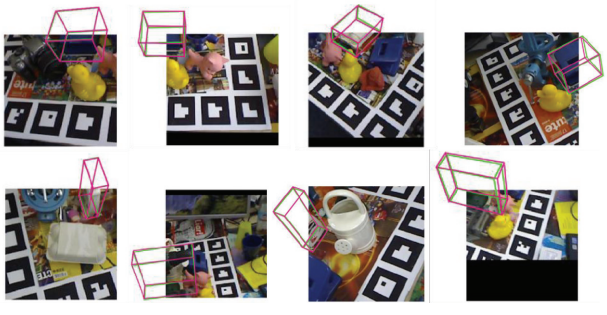


图 6 TruncationLINEMOD 数据集位姿可视化

Fig. 6 Visualization of poses on TruncationLINEMOD dataset

图 4 展示了 SC-Pose 对 LINEMOD 数据集图像中目标物体预测的位姿。图中,红色包围框表示 SC-Pose 预测的物体位姿,绿色包围框表示位姿真值。可以看出,SC-Pose 预测的物体位姿与真值非常接近。

图 5 展示了 SC-Pose 对 OCC-LINEMOD 数据集图像中目标物体预测的位姿。可以看出,SC-Pose 对于遮挡场景具有较好的位姿估计性能,即使目标物体一部分被其他物体遮挡,也可以准确地预测出物体位姿。

图 6 展示了 SC-Pose 对 TruncationLINEMOD 数据集图像中目标物体预测的位姿。可以看出,SC-Pose 对于截断场景具有良好的鲁棒性,即使目标物体区域大部分不可见,也可以准确地预测出物体位姿。

4 机器人抓取交互实验

为了评估 SC-Pose 在现实世界的机器人抓取任务中的性能,本文使用 RGB-D 相机 Realsense D415 采集场景图像,并使用带有三指夹爪的 Kinova Jaco2 机械臂进行机器人抓取交互实验。

为了更好地模拟实际抓取交互任务中物体尺寸大小不定的情况,本文选择了 LINEMOD 数据集中 5 个不同体积的物体模型(Glue、Ape、Duck、Cat 和 Phone)进行 3D 打印,并将其作为抓取交互实验物体。真实模型与 3D 打印模型的对比如图 7 所示。可以看出,3D 打印的物体模型与数据集中物体模型的颜色和纹理有一定的差异。

(a) 数据集中的物体模型 (b) 3D打印的物体模型

图 7 真实模型与 3D 打印模型的对比

Fig. 7 Comparison between real models and 3D printed models

机器人抓取交互实验中对每类交互物体设置 10 次抓取,每次抓取前随机确定不同的位置和角度来摆放交互物体,以机械臂末端手爪抓住物体且不掉落记为抓取成功,最后,统计总的抓取成功次数。

第 3 视角下机械臂抓取桌子上“Cat”物体的过程如图 8 所示。抓取成功率的统计结果见表 6。与采用 PVNet 模型的位姿估计结果相比,采用本文方法模型的位姿估计结果在小尺寸物体的抓取成功率提升明显且平均抓取成功率更高。

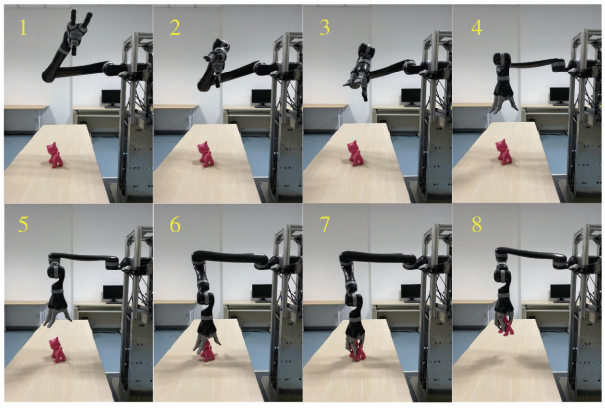


图 8 抓取“Cat”物体的过程

Fig. 8 Process of grasping the object “Cat”

表 6　单物体场景下的抓取成功率对比

Table 6　Comparison of grasping success rate in single object scene

位姿估计模型	交互物体（体积/cm^3）	抓取次数	成功次数	失败次数	成功率/%
PVNet[4]	Glue(494)	10	3	7	30
	Ape(540)	10	4	6	40
	Duck(689)	10	4	6	40
	Cat(1003)	10	4	6	40
	Phone(2553)	10	7	3	70
	平均	10	4.4	5.6	44.0
SC-Pose	Glue(494)	10	4	6	40
	Ape(540)	10	6	4	60
	Duck(689)	10	5	5	50
	Cat(1003)	10	7	3	70
	Phone(2553)	10	7	3	70
	平均	10	5.8	4.2	58.0

5　结论

1）本文提出 SC-Pose，通过约束预测关键点的拓扑结构与真实关键点的拓扑结构相似，使 2D 关键点的定位更加准确，最终提升 6D 位姿估计的精度。

2）本文在 6D 位姿估计常用数据集上进行了广泛实验，在 ADD(-S)指标的平均准确率方面优于主流的 6D 位姿估计方法，分别在 LINEMOD、OCC-LINEMOD 和 TruncationLINEMOD 数据集 ADD(-S)指标上取得了 90.90%、47.71% 和 35.39% 的平均准确率。

3）本文在真实场景中进行了机器人抓取交互实验，验证了 6D 位姿估计对机器人实施抓取交互的有效性。

参考文献：

[1] RAD M, LEPETIT V. BB8: a scalable, accurate, robust to partial occlusion method for predicting the 3D poses of challenging objects without using depth [C]//2017 IEEE International Conference on Computer Vision. Piscataway, NJ: 2017: 3848-3856.

[2] 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, NJ: 2018: 292-301.

[3] LEPETIT V, MORENO-NOGUER F, FUA P. EPnP: an accurate $O(n)$ solution to the PnP problem [J]. International Journal of Computer Vision, 2009, 81(2): 155-166.

[4] 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. Piscataway, NJ: 2019: 4561-4570.

[5] SONG C, SONG J R, HUANG Q X. HybridPose: 6D object pose estimation under hybrid representations [C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: , 2020: 431-440.

[6] XIONG F, LIU C J, CHEN Q J. Region pixel voting network (RPVNet) for 6D pose estimation from monocular image [J]. Applied Sciences, 2021, 11(2): 743.

[7] LIU X Y, IWASE S, KITANI K M. KDFNet: learning keypoint distance field for 6D object pose estimation [C] //2021 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, NJ: 2021: 4631-4638.

[8] YU X, ZHUANG Z Y, KONIUSZ P, et al. 6DoF object pose estimation via differentiable proxy voting regularizer [C] // British Machine Vision Conference. Berlin: Springer, 2020: 1-12.

[9] ZAKHAROV S, SHUGUROV I, ILIC S. DPOD: 6D pose object detector and refiner [C] // 2019 IEEE/CVF International Conference on Computer Vision. New York: IEEE Press, 2019: 1941-1950.

[10] PARK K, PATTEN T, VINCZE M. Pix2Pose: pixel-wise coordinate regression of objects for 6D pose estimation [C] // 2019 IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2019: 7668-7677.

[11] LIU X Y, JONSCHKOWSKI R, ANGELOVA A, et al. KeyPose: multi-view 3D labeling and keypoint estimation for transparent objects [C] // 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 11602-11610.

[12] HODAN T, BARATH D, MATAS J. EPOS: estimating 6D pose of objects with symmetries[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 11703-11712.

[13] XIANG Y, SCHMIDT T, NARAYANAN V, et al. PoseCNN: a convolutional neural network for 6D object pose estimation in cluttered scenes[EB/OL]. [2023-05-20]. http://export.arxiv.org/abs/1711.00199.

[14] 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, NJ: IEEE, 2017: 1521-1529.

[15] DO T T, CAI M, PHAM T, et al. Deep-6DPose: recovering 6D object pose from a single RGB image [EB/OL]. [2023-05-20]. https://arxiv.org/abs/1802.10367.

[16] 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. Piscataway, NJ: IEEE, 2019: 7678-7687.

[17] TIAN M, PAN L, ANG M H, et al. Robust 6D object pose estimation by learning RGB-D features [C] // 2020 IEEE International Conference on Robotics and Automation. Piscataway, NJ: IEEE, 2020: 6218-6224.

[18] WANG H, SRIDHAR S, HUANG J W, et al. Normalized object coordinate space for category-level 6D object pose and size estimation [C] // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 2642-2651.

[19] TIAN M, ANG M H Jr, LEE G H. Shape prior deformation for categorical 6D object pose and size estimation [C] // European Conference on Computer Vision. Cham: Springer, 2020: 530-546.

[20] PARK K, MOUSAVIAN A, XIANG Y, et al. LatentFusion: end-to-end differentiable reconstruction and rendering for unseen object pose estimation [C] // 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 10710-10719.

[21] CHEN D S, LI J, WANG Z, et al. Learning canonical shape space for category-level 6D object pose and size estimation [C] // 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 11973-11982.

[22] AMINI A, PERIYASAMY A S, BEHNKE S. T6D-Direct: transformers for Multi-object 6D pose direct regression [C] // DAGM German Conference on Pattern Recognition. Cham: Springer, 2021: 530-544.

[23] BUKSCHAT Y, VETTER M. EfficientPose: an efficient, accurate and scalable end-to-end 6D multi object pose estimation approach [EB/OL]. [2023-05-20]. https://arxiv.org/abs/2011.04307.

[24] SHEN W, ZHAO K, JIANG Y, et al. Object skeleton extraction in natural images by fusing scale-associated deep side outputs [C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 222-230.

[25] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770-778.

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

[27] BRACHMANN E, KRULL A, MICHEL F, et al. Learning 6D object pose estimation using 3D object coordinates [C] // European Conference on Computer Vision. Cham: Springer, 2014: 536-551.

[28] XIAO J X, HAYS J, EHINGER K A, et al. SUN database: large-scale scene recognition from abbey to zoo [C] // 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2010: 3485-3492.

(责任编辑 梁 洁)