控制与决策
Control and Decision

低质量渲染图像的目标物体6D姿态估计

左国玉， 张成威， 刘洪星， 龚道雄

引用本文:
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.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2020.1057

低质量渲染图像的目标物体6D姿态估计

左国玉†, 张成威, 刘洪星, 龚道雄

(1. 北京工业大学 信息学部，北京 100124；2. 北京市计算智能与智能系统重点实验室，北京 100124)

摘 要: 从图像中获取目标物体的6D位姿信息在机器人操作和虚拟现实等领域有着广泛的应用,然而,基于深度学习的位姿估计方法在训练模型时通常需要大量的训练数据集来提高模型的泛化能力,一般的数据采集方法存在收集成本高同时缺乏3D空间位置信息等问题. 鉴于此,提出一种低质量渲染图像的目标物体6D姿态估计网络框架. 该网络中,特征提取部分以单张RGB图像作为输入,用残差网络提取输入图像特征;位姿估计部分的目标物体分类流用于预测目标物体的类别,姿态回归流在3D空间中回归目标物体的旋转角度和平移矢量. 另外,采用域随机化方法以低收集成本方式构建大规模低质量渲染、带有物体3D空间位置信息的图像数据集Pose6DDR. 在所建立的Pose6DDR数据集和LineMod公共数据集上的测试结果表明了所提出位姿估计方法的优越性以及大规模数据集域随机化生成数据方法的有效性.

关键词: 6D位姿估计；域随机化；低质量渲染；RGB图像；Pose6DDR

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2020.1057 **开放科学(资源服务)标识码(OSID):**

引用格式: 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.

6D object pose estimation for low-quality rendering images

ZUO Guo-yu†, ZHANG Cheng-wei, LIU Hong-xing, GONG Dao-xiong

(1. Faculty of Information Technology，Beijing University of Technology，Beijing 100124，China；2. Beijing Key Laboratory of Computing Intelligence and Intelligent Systems，Beijing 100124，China)

Abstract: The 6D object pose obtained from single RGB image has broad applications such as robotic manipulation and virtual reality. However, the deep learning-based pose estimation methods usually require a large amount of trainging data to improve the generalization ability of the model, and in general, the common data generation methods have great challenges in high cost of data collection and lack of 3D information. This paper proposes a 6D object pose estimation network with low-quality rendering images. In this network, the feature extraction part takes a single RGB image as the input of the network, and uses the residual network to extract the features of this image. The classification stream of the pose estimation part predicts the category of the target object, and the regression stream returns the rotation angle and translation vector of the target object in 3D space. Moreover, the domain randomization method is used to establish a large-scale low-quality rendering images with the 3D spatial position information in a low collection cost. The experimental results on the established Pose6DDR dataset and the public LineMod dataset verify the superiority of the proposed pose estimation method and the effectiveness of the established large-scale simulation dataset.

Keywords: 6D pose estimation；domain randomization；low-quality rendering；RGB image；Pose6DDR

0 引 言

基于图像的6D目标姿态估计在虚拟现实和机器人操作等应用中发挥着越来越重要的作用. 在Amazon picking challenge[1]中,机器人需要从仓库货架中抓取对象物体,准确而鲁棒的姿态估计对于提高抓取的效率和成功率至关重要. 仅使用RGB图像估计位姿存在很多挑战,如物体检测时的严重遮挡、光线和外观的多样性以及背景物体的混乱状态. 传统方法通常会在3D模型与相应物体的2D图像之间建立对应关系,但是这些方法更多地依赖于手工制作模板的质量,对于具有杂乱背景的图像鲁棒性很差. 基于深度学习的方法以图像为输入,通过训练神经网络

收稿日期: 2020-07-30；**修回日期:** 2020-09-29.

基金项目: 国家重点研发计划项目(2018YFB1307004)；国家自然科学基金项目(61873008)；北京市自然科学基金项目(4182008, 4192010).

责任编委: 林崇.

†通讯作者. E-mail: zuoguoyu@bjut.edu.cn.

获取目标位姿估计,一般地,这种方法的泛化能力有待提高.

基于深度学习的位姿估计方法可以分一阶段估计方法和两阶段估计方法两类. 两阶段估计方法使用perspective-n-point(PnP)算法对得到的坐标进行解算,进而得到物体相对于相机的位姿[2-4]. 对于两阶段估计方法,已知目标物体的3D表面关键点,首先检测这些关键点在图像空间上的2D重投影坐标,得到2D关键点在图像中的位置;再通过PnP算法计算得到6D物体姿态估计参数. Rad等[5]通过预测目标物体的3D关键点在2D图像中的投影,采用PnP算法实现了目标物体的位姿估计. Tekin等[6]提出了一个更深层的卷积网络框架,该框架使用YOLOv2检测图像空间中8个边界框顶点的2D投影,再利用PnP算法计算目标物体的6D姿态信息. Peng等[7]提出了一种基于像素级投票网络架构的6D对象姿态估计方法,可以检测处于截断状态的目标对象的6D姿势. 两阶段方法的最终估计效果取决于2D关键点的检测精度,由于检测技术的发展,该方法可以取得非常高的估计精度. 但是两阶段方法很难处理遮挡和截断的问题,因为某些关键点会由于遮挡或截断而检测不到. 尽管卷积神经网络可以通过记忆估计模式来预测这些看不见的关键点,但提高泛化性仍然很困难.

提高系统的泛化性能可以通过增加网络的训练数据集来实现,但是当前的数据集规模普遍较小,有些方法可以获得大量的训练数据集,但又存在标签的标注以及收集成本过高的问题. 针对这些问题,可以通过仿真环境收集、图像变换、图像增强等方式收集用于训练的图像数据. 对仿真中学习到的模型进行微调要比在现实世界中从头学习来得更快[8-9]. 然而,真实环境中的光照条件和纹理等因素很难在仿真环境中完全重现,仿真图像与真实图像之间存在较大差距;想要通过使用仿真图像代替真实图像来训练模型就要克服这些问题. 减少仿真图像与真实图像数据之间差异的方法有多种,如文献[10]使用高保真仿真环境渲染图片,使用高仿真合成图像进行目标评估,但是这种方法在复杂场景中的性能较差,同时需要大量的计算资源和高质量的模型. 域随机化方法是一种减少仿真图像和真实图像数据之间差异的方法,Tobin等[11]使用该方法生成模型以消除其间的差异,并使用仿真图像训练自回归模型来学习目标物体的位姿信息,用于机械臂抓取任务.

一阶段估计方法无需利用PnP算法,它可以基于回归直接从输入图像获取目标物体6D位姿估计参数. Xiang等[12]提出了6D对象姿态估计模型PoseCNN,使用霍夫投票确定对象位置的中心,预测与摄像机的距离,进而估计目标的3D平移,并通过返回四元数计算3D旋转,PoseCNN提出了新的损失函数Shape Match-Loss,用于旋转对称对象的姿态估计. Do等[13]提出了一种Deep-6DPose端到端深度学习框架,可以从单个RGB图像中检测、分割和恢复对象实例的6D姿态信息.

本文采用一阶段估计方法从单个RGB图像中获取物体姿态. 不同于其他一阶段方法,该方法通过直接回归得到目标物体的旋转角度和平移矢量,以实现目标物体的6D位姿估计. 同时,针对训练数据集不足的情况,将模拟数据扩展到6D姿态估计领域,并使用域随机化方法生成大规模的数据集图像用于目标物体6D姿态估计,通过生成的模拟数据提高6D姿态估计的准确性和泛化性能. 使用Pose6DDR数据集和LineMod[14]数据集对所提出的网络进行训练,并在这两个数据集上对所提出的网络结构以及基于域随机化建立的数据集对于提高目标物体6D姿态估计性能的有效性进行了验证.

1 基于域随机化的数据集建立

下面详细介绍Pose6DDR数据集的建立方法,包括仿真环境、对象模型的建立和生成、对象标签注释以及仿真环境中的随机化因素. 数据集共收集超过80 000张低质量渲染的仿真图像,且与LineMod数据集图像具有高度的相似性. 通过实验验证了该数据集能够提高目标物体的位姿估计精度,由于仿真图像在变换随机因素的场景下生成,与仅使用LineMod图像训练的网络相比,使用生成的仿真图像训练的网络具有更好的泛化能力.

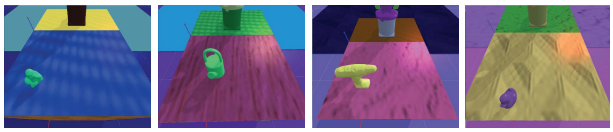
1.1 环境和模型生成

基于Bullet物理引擎搭建与现实环境比例相同的外部图像收集环境,同时收集包括猿猴、浇水壶、猫、钻、鸭子、鸡蛋盒、胶水和打孔器8种目标物体的模型数据,使环境中建立的对象模型大小比例与公共数据集LinMod的模型比例相同. 使用域随机化方法减小仿真图像与真实图像之间的差距. 在图像收集过程中,主要思想是通过以低质量渲染更改图像空间中的因素生成Pose6DDR数据集. 该方法随机生成工作台和操作机械臂的纹理、位置、光照强度、光照方向等因素,同时在仿真环境中匹配LineMod数据集中的对象模型分布,因此收集的仿真图像更加接近真实图像,如图1所示. 与LineMod拥有1 213×8张图像相比,Pose6DDR数据集拥有超过80 000张图像,而收集

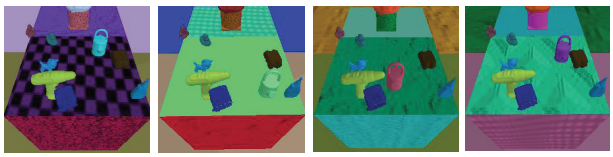
成本相对很低,在一台拥有GTX 1080Ti GPU的电脑上运行两天即可完成收集工作. 由图1可见,合成的数据图像与LineMod数据集图像相似. 图1所示图像在以下两个方面进行了随机:

1) 内容变化. 在所建立的外部环境中放入目标物体的3D模型,这些模型以随机的位置和朝向被放置在所建立的工作台上. 为了获得类似于LineMod数据集的图像,使用超过8种不同类型的对象模型,这些对象模型与LineMod数据集的模型大小比例一致. 通过固定工作台上除目标物体以外其他类型的对象模型,使得模型分布与LineMod数据集中模型分布相似.

2) 样式变化. 样式变化是指随机更改所有对象的颜色、纹理和光照等因素. 本文通过随机产生目标物体的颜色、纹理、位置和旋转角度等因素来收集图像,具有不同纹理的纹理库用于实现样式变化,应用光增强捕获变化的阴影条件和时间变化,通过改变光源的位置和方向实现照明条件的改变.



(a) 对象模型



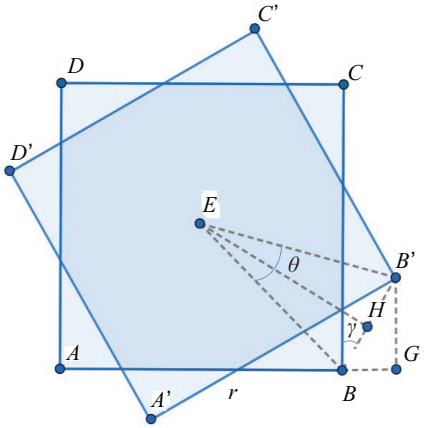
(b) Pose6DDR 数据集

(c) 公共数据集 LineMod

图 1　数据集图像

1.2　标签注释

一般而言,仿真数据集作为训练数据会存在数据集标签制作的难度. 本文Pose6DDR数据集用于训练位姿估计网络,在生成该数据集的过程中,针对任务目标对图像中目标物体进行标签注释. 为获取目标物体的类别,对于物体的横滚、俯仰、偏航的旋转角度以及对象的平移矢量,Pose6DDR设有16位注释标签,其中包含类别(1位)、3个角度(3×1位)、旋转矩阵(9位)和平移矩阵(3位),分类任务标签由one-hot编码表示. 获取数据的标签信息需要获得目标物体3D空间中边界框的8个顶点坐标,即图2所示目标物体各个顶点坐标.



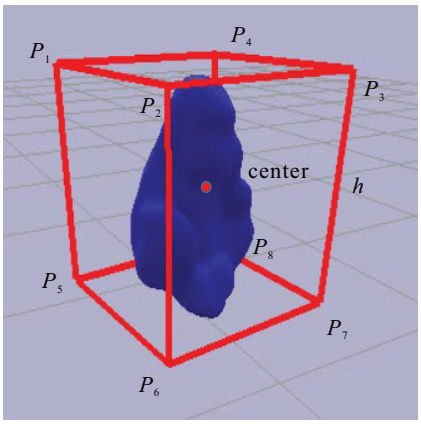


图 2　目标物体坐标示意图

正方形$ABCD$是旋转之前3D边界框的俯视图,正方形$A'B'C'D'$是旋转了一定角度$\theta = \angle B'EB$的边界框的俯视图,r是3D边界框的底边长度,h是3D边界框的高度. 根据图示几何关系,有

$$EB = \frac{\sqrt{2}}{2}r, \tag{1}$$

$$BH = EB \times \sin\frac{\theta}{2}, \tag{2}$$

$$BB' = 2BH = 2\frac{\sqrt{2}}{2}r \times \sin\frac{\theta}{2}, \tag{3}$$

$$\angle EBB' = \angle BB'G = \frac{180 - \theta}{2} - 45. \tag{4}$$

令$l = BB', \gamma = \angle CBB' = \angle BB'G, B$的坐标为$(x_B, y_B)$,可以计算得到点$B'$的坐标为$(x_B - \cos\gamma \times l, y_B + \sin\gamma \times l)$. 目标对象的中心点与3D边界框的中心重合,每个点在空间中的位置如图2所示. 目标对象中心点的坐标为(x, y, z),3D边界框的8个顶点的坐标可以分别通过类似的转换获得.

2　姿态估计网络框架

利用深度学习方法构建一个对目标物体的6D姿态进行估计的网络框架. 通过输入单张RGB图像,输出目标物体的对象姿态估计,进而实现机器人抓取任务,其中输入的大规模RGB图像数据集通过前文域随机化方法生成.

2.1 6D位姿表示

由单张RGB图像直接回归得到目标物体6D位姿存在很多困难,如基于单张RGB图像的6D姿态估计缺少3D模型信息,需要进行后处理等. 当RGB图像作为网络的输入时,由于没有完整的3D空间表示形式,需要先进行位姿信息的表示. 本文方法平移矩阵由网络的输出表示,旋转矩阵的获取比平移矩阵更复杂,要通过旋转角度计算得到. 由于旋转角以2π弧度为一个周期,相同的角度可以用多个值表示,这里以获得的最小值为旋转角度. 将旋转角度和平移向量用于计算目标物体相对于相机的旋转矩阵和平移矩阵,有

$$R_x(\varphi)=\begin{bmatrix}1 & 0 & 0\\ 0 & \cos\varphi & \sin\varphi\\ 0 & -\sin\varphi & \cos\varphi\end{bmatrix}, \tag{5}$$

$$R_y(\theta)=\begin{bmatrix}\cos\theta & 0 & -\sin\theta\\ 0 & 1 & 0\\ \sin\theta & 0 & \cos\theta\end{bmatrix}, \tag{6}$$

$$R_z(\psi)=\begin{bmatrix}\cos\psi & \sin\psi & 0\\ -\sin\psi & \cos\psi & 0\\ 0 & 0 & 1\end{bmatrix}, \tag{7}$$

其中的φ、θ、ψ分别为目标物体的横滚、俯仰和偏航角度. 3D旋转矩阵表示为$R=R_x\cdot R_y\cdot R_z$.

2.2 网络结构

图3为所提出的位姿估计网络框架,该网络包括特征提取和姿态估计两个部分. 特征提取部分以残差网络为骨架,用于提取输入图像的特征并传输给位姿估计部分. 位姿估计部分包含两个分支,一是目标物体分类流,二是目标物体姿态回归流. 分类流用于预测目标物体的类别,姿态回归流在3D空间中回归目标物体的旋转角度和平移矢量,进而计算得到相对于相机的旋转矩阵和平移矩阵.

2.3 多目标损失函数

进行网络训练时,需要设计针对多目标多任务的损失函数来训练分类任务、旋转角度和平移矢量任务. 损失函数总体表示如下:

$$\ell=\alpha_1\ell_{\rm cls}+\alpha_2\ell_{\rm pose}. \tag{8}$$

其中: 分类损失$\ell_{\rm cls}$为softmax损失函数,$\ell_{\rm pose}$为位姿损失函数,α_1、α_2为控制训练过程中每项损失重要性的比例因子. 姿态估计流输出6个向量分别有3个旋转角度和3个平移矢量. 位姿损失$\ell_{\rm pose}$定义为

$$\ell_{\rm pose}=\|r-\hat{r}\|+\beta\|t-\hat{t}\|. \tag{9}$$

其中: r为回归得到的旋转角度,$\hat{r}$为旋转角度的标签真值,t为回归得到的平移矢量,$\hat{t}$为平移矢量的标签真值数据,β为控制旋转和平移回归误差的比例因子.

2.4 训练和测试

使用PyTorch框架搭建网络结构,在训练和测试过程中,以单个RGB图像作为输入,网络的输出包括目标对象的类别、目标对象的3个旋转向量和3个平移向量,用于计算目标物体的旋转矩阵和平移矩阵. 在训练过程中,式(8)和(9)中的α_1、α_2和β经实验

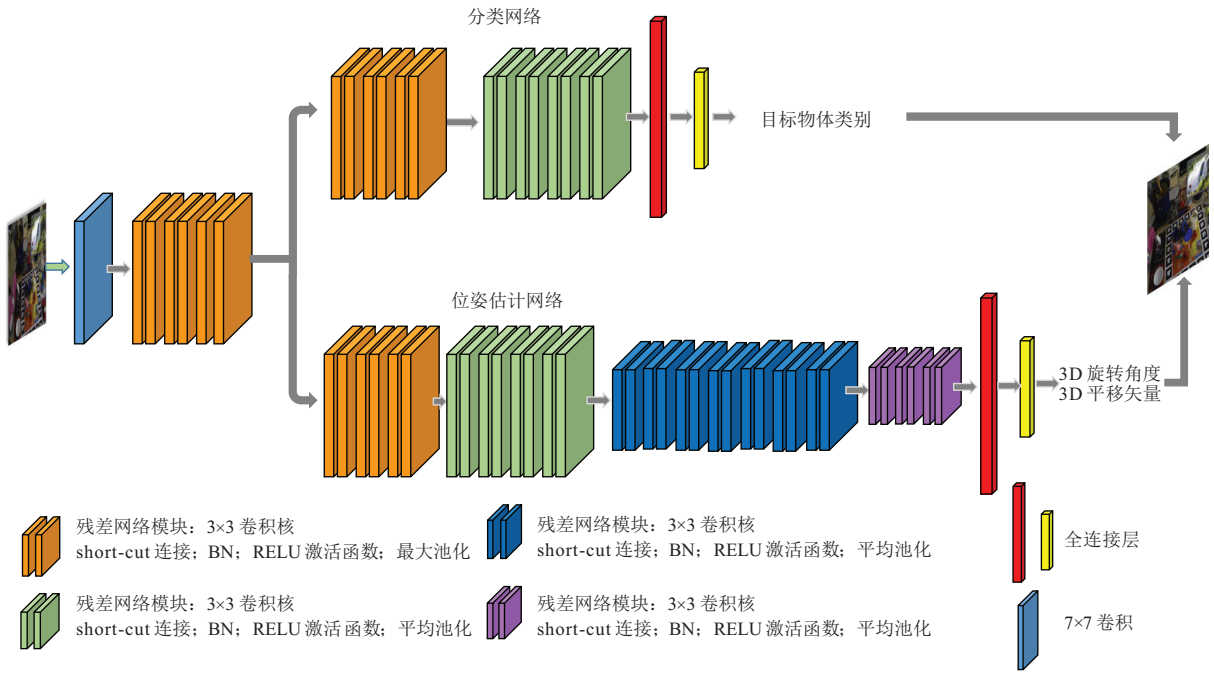


图3 网络结构

测试分别设置为0.7、1和0.7. 在训练的反向传递过程中,由于任务更加侧重于目标物体的位姿估计,通过设置不同的比例参数以提高位姿估计误差的作用,这里将ℓ_{pose}的权重参数设置为1可以提高优化性能. 该网络结构在GTX 1080Ti GPU上进行200个epochs,训练使用momentum为0.1的Adam优化器,权重系数为0.000 1,每个小批量输入图像有128张图像,前20个epochs的学习率设置为0.000 1,之后每20个epochs学习率减小10倍.

3　实验和分析

为了评估所提出方法的有效性,使用Pose6DDR数据集和LineMod数据集对位姿估计网络进行训练和测试,公共数据集LineMod是单个RGB目标图像数据集. 在LineMod数据集中,每个RGB图像都有目标物体的旋转矩阵、平移矩阵和ID的真实标签数据. 该数据集有13个对象和3D模型,每个对象有约1 213张图像. Pose6DDR数据集具有超过80 000张RGB图像.

为了方便对比实验效果,使用通用的量化指标5 cm 5°和ADD指标,通过计算网络预测输出的参数与数据标签真值获得的姿态之间的误差评估方法性能. 采用5 cm 5°指标时,与真实标签数据相比,如果平移误差和角度误差分别在5 cm 和5°范围内,则认为该预测输出符合指标,接受该值为正确预测值. 采用ADD时,如果位姿估计与变换后模型点云之间真实姿态的平均距离小于对象直径的10 %,则接受该预测估计位姿. ADD指标计算如下:

$$s=\frac{1}{|M|}\sum_{x_1\in M}\min_{M}\|(Rx+t)-(\hat{R}x+\hat{t})\|. \tag{10}$$

其中: R和t为真实标签数据,$\hat{R}$和$\hat{t}$为网络预测的旋转和平移矩阵,M为3D模型的顶点坐标集.

3.1　Pose6DDR数据集的测试结果

在Pose6DDR数据集上测试该网络结构的有效性. 数据集分为训练和测试两部分,部分RGB图像用于训练网络,另2 000张合成图像用于测试. 除5 cm 5°外,其他度量标准通过减小角度和距离来测试网络结构的鲁棒性. 表1展示了具有不同度量标准的Pose6DDR数据集的测试结果(单位: %).

表1中: 第1列为5 cm 5°测试结果,可以看出,除了鸡蛋盒74.8 %和鸭子93.3 %外,其他几类物体的准确性均超过95 %. 后4列分别为将指标提高至3 cm 5°、1 cm 5°、5 cm 3°和5 cm 1°的结果,可以看出,当度量标准改变时,对象的角度对精度结果的影响更大,当预测的旋转角度在3° ~ 5°之间时,钻的结果均为零,表明网络的鲁棒性在角度指标上效果较差,而距离指标相对于角度指标鲁棒性更好. 图4为合成图像上目标物体估计结果的可视化示例.

表 1　Pose6DDR数据集在5 cm 5°指标下的测试结果

目标物体	5 cm 5°	3 cm 5°	1 cm 5°	5 cm 3°	5 cm 1°
猿猴	**96.9**	89.1	64.0	4.1	0
浇水壶	**98.3**	77.6	33.6	**98.3**	2.9
猫	**98.5**	93.7	83.9	82.1	0
钻	**99.6**	97.4	94.5	0	0
鸭子	**93.3**	76.4	47.5	48.2	1.2
鸡蛋盒	**74.8**	52.2	25.2	60.0	1.3
胶水	**96.2**	90.2	65.9	15.0	0.2
打孔器	**96.1**	78.8	62.8	**96.1**	0
平均值	**94.2**	81.9	59.7	50.5	0.7

图 4　Pose6DDR图像可视化结果

3.2　消融实验

通过消融实验分析域随机化中每个因素对最终结果的影响. 实验中,每次固定一个随机化因素生成数据集,使用该数据集训练并测试网络模型,观察其对最终预测的影响. 实验分别收集固定纹理、固定光照方向、不进行光增强3个因素的数据集. 为了保证结果不受数据集规模大小的影响,收集的数据集图像数量与完全域随机化数据集的数量一致,每类随机因素均收集超过80 000张仿真图像. 使用数据集对网络进行训练,将其与完全域随机化生成的图像测试结果进行对比,如表2所示(单位: %).

表 2　不同随机因素在5 cm 5°指标下的测试结果

目标物体	固定纹理	固定光照(方向)	无光增强	完全随机
猿猴	95.8	94.3	91.6	**96.9**
浇水壶	96.3	93.6	91.1	**98.3**
猫	97.9	94.8	93.5	**98.5**
钻	98.1	94.9	92.8	**99.6**
鸭子	91.6	89.8	87.5	**93.3**
鸡蛋盒	70.6	68.9	64.5	**74.8**
胶水	94.5	92.8	90.6	**96.1**
打孔器	95.3	92.6	89.7	**96.1**
平均值	92.5	90.2	87.7	**94.2**

完全由域随机化得到的仿真图像目标物体的位姿预测精度最高. 对于纹理、光照和光增强3种随机因素,固定其中任何一个都会对生成的图像产生影响,并降低最终的估计精度. 但是,通过对比光照和纹理的实验结果可以看到,固定光照和无光增强的准确率更低,表明光照对整体图像的生成质量有非常大的影响,从而降低了最终的估计精度. 分析其原因,在相机收集图像的过程中,物体表面会由于光照方向的不同和光照强度的不同产生不同强度的光反射,影响目标物体生成的质量,从而使收集的图像成像质量产生较大波动,最终在进行位姿估计时其精度大幅下降. 纹理相对于光照而言,对图像的生成质量影响较小,最终的估计精度虽然会下降,但是下降幅度很小. 由此可以推出结论,在收集仿真图像时,对图像的成像质量有较大影响的因素进行随机非常必要,完全随机化对数据集的质量以及对提高最终任务的估计精度均有非常重要的作用.

3.3 LineMod数据集上的结果分析

为了评估所提出方法的优越性,实验使用LineMod数据集对其进行测试,并与其他采用5 cm 5°和ADD指标的方法进行比较. 实验中,对于每一类物体都有超过200张图像用于测试. 由于本文方法没有对结果进行后处理,将本文方法与没有进行后处理的当前流行方法进行对比,包括Brachmann[15]、BB8[5]和Deep6DPose[13]. 表3和表4为两个指标下的位姿估计结果(单位:%).

表3 LineMod数据集在5 cm 5°指标下的测试结果

目标物体	猿猴	浇水壶	猫	钻	鸭子	鸡蛋盒	胶水	打孔器	平均值
Deep6DPose	57.8	70.1	70.3	72.9	67.1	68.4	64.6	70.4	67.7
Branchmann	34.4	48.4	34.6	54.5	22.0	57.1	23.6	47.3	40.2
BB8	80.2	76.8	79.9	69.6	53.2	81.3	54.0	73.1	71.0
本文方法	57.5	63.0	45.0	54.0	56.5	**92.5**	38.0	47.5	56.8

表4 LineMod数据集在ADD指标下的测试结果

目标物体	猿猴	浇水壶	猫	钻	鸭子	鸡蛋盒	胶水	打孔器	平均值
Deep6DPose	38.8	86.1	66.2	82.3	32.5	79.4	63.7	56.4	63.2
Branchmann	33.2	62.9	42.7	61.9	30.2	49.9	31.2	52.8	45.6
BB8	27.9	48.1	45.2	58.6	32.8	40.0	27.0	42.4	40.3
本文方法	49.5	73.0	51.0	42.0	49.0	**79.5**	52.5	47.0	55.4

由表3中5 cm 5°测试结果可见,BB8的8个对象平均准确率比Deep6DPose高约3.3%,Deep6DPose比Brachmann高27.5%,Deep6DPose相比于BB8和Brachmann更加稳定. 值得注意的是,本文的5 cm 5°结果平均值比Deep6DPose稍低一些,但比Brachmann高12.3%. 此外,在本文方法的结果中,某些对象的精度高于Deep6DPose和BB8,如鸡蛋盒的精度为92.5%. 本文方法在鸭子和鸡蛋盒两类物体的准确率上都超过了BB8. 相对而言,本文方法在胶水对象目标上的效果最差,其预测平移与标签真值数据之间的差远远超过了度量指标. 与Deep6DPose相比,可以看出本文方法在某些物体上还不够稳定.

表4为使用ADD度量指标下8个对象的位姿估计结果. 可以看到,Deep6DPose方法的平均准确性比BB8高22.9%,本文方法在BB8和Brachmann方面的平均表现较好,但略低于Deep6DPose. 鸭子在所有方法中的准确性都较低,但本文方法为49.0%比其他方法要高. 总体而言,本文方法获得了具有竞争力的准确率结果,在ADD度量下的结果比其他方法更稳定. 平均而言,尽管所提出方法并不是最好,但它已胜过一些方法,并在目标物体猿猴、鸭子和鸡蛋盒上获得了最高的估计精度. 实验中发现,使用LineMod数据集训练该网络结构比使用Pose6DDR数据集要花费更多的时间,且后者具有更高的性能. 图5为一些LineMod数据集上单个目标物体对象位姿估计的结果示例.

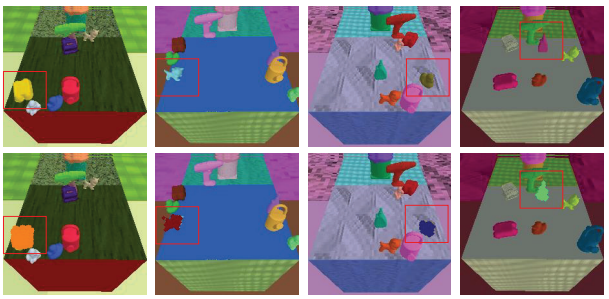


图5 LineMod数据集图像可视化结果

4 结 论

本文提出了一种用于从单张RGB图像中获得目标物体6D位姿的网络结构. 该网络结构可以直接输出估计的旋转角度和平移矢量,并计算得到相对于相机的6D位姿参数. 此外,采用低质量驱动的渲染技术建立Pose6DDR大规模仿真数据集,该数据集具有超过80 000张图像,且收集成本低. 通过仿真数据集训练6D姿态估计网络,表明了数据集的有效性. 与其他基于RGB图像的6D物体位姿估计方法的对比结果显示,所提出方法在某些物体上具有明显的准确率提升,显示出明显的优越性. 接下来的工作,将使用不同的方法扩展数据集,并改善网络处理更加复杂场景图像时的性能.

参考文献(References)

[1] Correll N, Bekris K E, Berenson D, et al. Analysis and observations from the first Amazon picking challenge[J]. IEEE Transactions on Automation Science and Engineering, 2016, 15(1): 172-188.

[2] Lepetit V, Fua P. Monocular model-based 3D tracking of rigid objects: A survey[J]. Foundations and Trends in Computer Graphics and Vision, 2005, 1(1): 1-89.

[3] Rothganger F, Lazebnik S, Schmid C, et al. 3D object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints[J]. International Journal of Computer Vision, 2006, 66(3): 231-259.

[4] Wagner D, Reitmayr G, Mulloni A, et al. Pose tracking from natural features on mobile phones[C]. The 7th IEEE/ACM International Symposium on Mixed and Augmented Reality. Cambridge, 2008: 125-134.

[5] 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]. IEEE International Conference on Computer Vision. Venice, 2017: 3848-3856.

[6] Tekin B, Sinha S N, Fua P. Real-time seamless single shot 6D object pose prediction[C]. IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 292-301.

[7] Peng S D, Liu Y, Huang Q X, et al. PVNet: Pixel-wise voting network for 6DoF pose estimation[C]. IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, 2019: 4556-4565.

[8] Cutler M, How J P. Efficient reinforcement learning for robots using informative simulated priors[C]. International Conference on Robotics and Automation. Seattle, 2015: 2605-2612.

[9] Kolter J Z, Ng A Y. Learning omnidirectional path following using dimensionality reduction[C]. Robotics: Science and Systems III. Robotics: Science and Systems Foundation, 2007: 27-30.

[10] Movshovitz-Attias Y, Kanade T, Sheih Y. How useful is photo-realistic rendering for visual learning?[C]. European Conference on Computer Vision. Cham: IEEE, 2016: 202-217.

[11] Tobin J, Biewald L, Duan R, et al. Domain randomization and generative models for robotic grasping[C]. International Conference on Intelligent Robots and Systems. Madrid, 2018: 3482-3489.

[12] Xiang Y, Schmidt T, Narayanan V, et al. Posecnn: A convolutional neural network for 6D object pose estimation in cluttered scenes[J]. 2017, arXiv: 1711.00199.

[13] Do Thanh-Toan, Cai Ming, Pham T, et al. Deep-6dpose: Recovering 6d object pose from a single rgb image[J]. 2018, arXiv: 1802.10367.

[14] Hinterstoisser S, Lepetit V, UiC S, et al. Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes[J]. Computer Vision-ACCV, DOI: 10.1007/978-3-642-37331-2_42.

[15] Brachmann E, Michel F, Krull A, et al. Uncertainty-driven 6D pose estimation of objects and scenes from a single RGB image[C]. IEEE Conference on Computer Vision and Pattern Recognition. Berlin: IEEE, 2016: 3364-3372.

作者简介

左国玉 (1971–), 男, 教授, 博士, 从事机器人控制、机器人学习、智能计算等研究, E-mail: zuoguoyu@bjut.edu.cn;

张成威 (1996–), 男, 硕士生, 从事机器视觉、深度学习的研究, E-mail: ZCW0356@emails.bjut.edu.cn;

刘洪星 (1996–), 男, 硕士生, 从事机器视觉、机器人认知的研究, E-mail: xingl@emails.bjut.edu.cn;

龚道雄 (1968–), 男, 教授, 从事进化计算、智能机器人控制等研究, E-mail: gongdx@bjut.edu.cn.

(责任编辑: 郑晓蕾)