**PoseCNN：一种用于混乱场景中的6D目标姿态估计的卷积神经网络**

**PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes**

**——Yu Xiang、Tanner Schmidt、Venkatraman Narayanan、Dieter Fox (arXiv 2018)**

**一、科学问题**

**1.1 本文所涉及科学问题**

本文研究了物体的多样性以及由物体之间的杂乱和遮挡引起场景的复杂性的6D对象姿态估计问题。

**1.2 同行专家如何解决**

6D对象姿态估计方法可大致分为基于模板的方法和基于特征的方法：

1. 在基于模板的方法中，构造刚性模板并用于扫描输入图像中的不同位置。
2. 在基于特征的方法中，从图像中的每个像素点提取局部特征，并且与3D模型上的特征匹配怡建立2D-3D对应关系，从中可以恢复6D姿态。

**1.3 本文所解决的问题**

本文证明了PoseCNN对遮挡具有很强的鲁棒性，能够处理对称对象，并且仅使用彩色图像作为输入就能提供精确的姿态估计。当使用深度数据进一步细化姿态时，PoseCNN在OccludedLINEMOD数据集上获得最优性能。

**1.4 本文解决方案效果**

即使对象中心被另一个对象遮挡，对象中心的预测也是相当精确。PoseCNN网络对于彩色图像已经能够提供良好的6D姿态估计，通过ICP细化进一步提高了6D姿态的精度。

**二、研究内容**

**2.1 理论与方法介绍**

PoseCNN网络包含两个阶段。第一阶段由13个卷积层和4个最大池化层组成，它们从输入图像中提取具有不同分辨率的特征图。第二阶段包含一个嵌入步骤，该步骤将第一阶段生成的高维特征图嵌入到低维、特定任务的特征中，然后网络执行6D姿态估计的三个任务：语义标记，3D平移估计和3D旋转回归。

**2.2 验证分析与实验效果**

1. 通过仅使用彩色图像，PoseCNN网络在6D姿态估计方面明显优于3D坐标回归网络和RANSAC算法。
2. 采用ICP对姿态进行细化显著提高了性能。
3. 3D坐标回归网络不能很好地处理对称对象。

**三、论文存在问题及后续研究重点**

**3.1 论文存在问题**

在分析旋转回归的两个损失函数对对称对象的影响时，YCB-Video数据集中的两个对称对象存在0°和180°的旋转误差，因为它们相对于围绕其坐标轴的180°旋转是对称的。

**3.2 后续研究重点**

SLOSS有时会导致姿态空间中的局部最小值与ICP相似，在未来的6D姿态估计中探索处理对称对象的更有效方法.

**四、该问题相关研究成果**

**4.1 相关论文一**

**（1）题目**：Learning 6D Object Pose Estimation Using 3D Object Coordinates

**（2）作者介绍**：Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, Carsten Rother

**（3）摘要**: This work addresses the problem of estimating the 6D Pose of specific objects from a single RGB-D image. We present a flexible approach that can deal with generic objects, both textured and texture-less. The key new concept is a learned, intermediate representation in form of a dense 3D object coordinate labelling paired with a dense class labelling. We are able to show that for a common dataset with texture-less objects, where template-based techniques are suitable and state of the art, our approach is slightly superior in terms of accuracy. We also demonstrate the benefits of our approach, compared to template-based techniques, in terms of robustness with respect to varying lighting conditions. Towards this end, we contribute a new ground truth dataset with 10k images of 20 objects captured each under three different lighting conditions. We demonstrate that our approach scales well with the number of objects and has capabilities to run fast.

**4.2 相关论文二**

**（1）题目**：Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes

**（2）作者介绍**：Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, Nassir Navab

**（3）摘要**：We propose a framework for automatic modeling, detection, and tracking of 3D objects with a Kinect. The detection part is mainly based on the recent template-based LINEMOD approach for object detection. We show how to build the templates automatically from 3D models, and how to estimate the 6 degrees-of-freedom pose accurately and in real-time. The pose estimation and the color information allow us to check the detection hypotheses and improves the correct detection rate by 13% with respect to the original LINEMOD. These many improvements make our framework suitable for object manipulation in Robotics applications. Moreover we propose a new dataset made of 15 registered, 1100+ frame video sequences of 15 various objects for the evaluation of future competing methods.

**4.3 相关论文三**

**（1）题目**：Multi-view self-supervised deep learning for 6D pose estimation in the Amazon Picking Challenge

**（2）作者介绍**：Andy Zeng, Kuan-Ting Yu, Shuran Song, Daniel Suo, Ed Walker, Alberto Rodriguez, Jianxiong Xiao

**（3）摘要**：Robot warehouse automation has attracted significant interest in recent years, perhaps most visibly in the Amazon Picking Challenge (APC) [1]. A fully autonomous warehouse pick-and-place system requires robust vision that reliably recognizes and locates objects amid cluttered environments, self-occlusions, sensor noise, and a large variety of objects. In this paper we present an approach that leverages multiview RGB-D data and self-supervised, data-driven learning to overcome those difficulties. The approach was part of the MIT-Princeton Team system that took 3rd- and 4th-place in the stowing and picking tasks, respectively at APC 2016. In the proposed approach, we segment and label multiple views of a scene with a fully convolutional neural network, and then fit pre-scanned 3D object models to the resulting segmentation to get the 6D object pose. Training a deep neural network for segmentation typically requires a large amount of training data. We propose a self-supervised method to generate a large labeled dataset without tedious manual segmentation. We demonstrate that our system can reliably estimate the 6D pose of objects under a variety of scenarios. All code, data, and benchmarks are available at <http://apc.cs.princeton.edu/>.

**PVNet：用于六自由度姿态估计的像素投票网络**

**PVNet: Pixel-wise Voting Network for 6DoF Pose Estimation**

**——Sida Peng、Yuan Liu、Qixing Huang、Hujun Bao、Xiaowei Zhou (arXiv 2018)**

**一、科学问题**

**1.1 本文所涉及科学问题**

本文研究了单张RGB图像在严重遮挡或截断的条件下六自由姿态估计的问题。

**1.2 同行专家如何解决**

1. 可以通过建立对象图像和对象模型之间的对应关系来实现姿态估计
2. 首先使用CNN回归2D关键点，然后使用Perspective-n-Point(PnP)算法计算6D姿态参数。

**1.3 本文所解决的问题**

本文引入了一种新的六自由度对象姿态估计框架，它包括用于关键点定位的像素投票网络（PVNet）和用于最终姿态估计的不确定性驱动PnP。

**1.4 本文解决方案效果**

1. 发现通过基于RANSAC的关键点定位投票来预测矢量场比关键点坐标的直接回归获得了更好的性能，特别是对于被遮挡或截断的对象。
2. 在求解PnP问题时考虑了预测关键点位置的不确定性，进一步改进了姿态估计。
3. 实现了三种广泛使用的基准数据集的最优性能，并在一个新的截断对象数据集上证明了该方法的鲁棒性。

**二、研究内容**

**2.1 理论与方法介绍**

1. 使用两级管道估计对象姿态：首一类似先使用CNN检测2D对象关键点，然后使用PnP算法计算6D姿态参数。
2. 使用像素投票网络（PVNet）以类似RANSAC的方式检测2D关键点。
3. 基于RANSAC的投票给出了每个关键点的空间概率分布，用不确定性驱动的PnP估计6D姿态。

**2.2 验证分析与实验效果**

1. 本方法在LINEMOD数据集的所有对象上实现了最优性能，而不需要单独的细化阶段。与未使用细化的方法相比，性能至少提高30.32%，与使用边缘对齐来细化的SSD-6D相比，性能至少提高7.27%。
2. 矢量场表示方法使PVNet能够学习对象各部分之间的关系，使得被遮挡的关键点和姿态能够被可见部分鲁棒地恢复。
3. 性能优于专门为处理遮挡的Oberweger方法。

**三、论文存在问题及后续研究重点**

**3.1 论文存在问题**

对B.Tekin, S.N.Sinha, and P.Fua. Real-time seamless single shot 6d obejet pose prediction. In CVPR,2018.发布的模型进行了2D投影和ADD(-S)度量测试，但是由于不是针对这种情况设计的，所以没有得到合理的结果。

**3.2 后续研究重点**

对可见部分模糊的对象进行姿态估计。

**四、该问题相关研究成果**

**4.1 相关论文一**

**（1）题目**：Real-Time Seamless Single Shot 6D Object Pose Prediction

**（2）作者介绍**：Bugra Tekin, Sudipta N. Sinha, Pascal Fua

**（3）摘要**: We propose a single-shot approach for simultaneously detecting an object in an RGB image and predicting its 6D pose without requiring multiple stages or having to examine multiple hypotheses. Unlike a recently proposed single-shot technique for this task [Kehl et al. 2017] that only predicts an approximate 6D pose that must then be refined, ours is accurate enough not to require additional post-processing. As a result, it is much faster - 50 fps on a Titan X (Pascal) GPU - and more suitable for real-time processing. The key component of our method is a new CNN architecture inspired by [Redmon et al. 2016, Redmon and Farhadi 2017] that directly predicts the 2D image locations of the projected vertices of the object's 3D bounding box. The object's 6D pose is then estimated using a PnP algorithm. For single object and multiple object pose estimation on the LineMod and Occlusion datasets, our approach substantially outperforms other recent CNN-based approaches [Kehl et al. 2017, Rad and Lepetit 2017] when they are all used without post-processing. During post-processing, a pose refinement step can be used to boost the accuracy of these two methods, but at 10 fps or less, they are much slower than our method.

**4.2 相关论文二**

**（1）题目**：6-DoF object pose from semantic keypoints

**（2）作者介绍**：Georgios Pavlakos, Xiaowei Zhou, Aaron Chan, Konstantinos G. Derpanis, Kostas Daniilidis

**（3）摘要**：This paper presents a novel approach to estimating the continuous six degree of freedom (6-DoF) pose (3D translation and rotation) of an object from a single RGB image. The approach combines semantic keypoints predicted by a convolutional network (convnet) with a deformable shape model. Unlike prior work, we are agnostic to whether the object is textured or textureless, as the convnet learns the optimal representation from the available training image data. Furthermore, the approach can be applied to instance- and class-based pose recovery. Empirically, we show that the proposed approach can accurately recover the 6-DoF object pose for both instance- and class-based scenarios with a cluttered background. For class-based object pose estimation, state-of-the-art accuracy is shown on the large-scale PASCAL3D+ dataset.

**4.3 相关论文三**

**（1）题目**：SSD-6D: Making RGB-Based 3D Detection and 6D Pose Estimation Great Again

**（2）作者介绍**：Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, Nassir Navab

**（3）摘要**：We present a novel method for detecting 3D model instances and estimating their 6D poses from RGB data in a single shot. To this end, we extend the popular SSD paradigm to cover the full 6D pose space and train on synthetic model data only. Our approach competes or surpasses current state-of-the-art methods that leverage RGB-D data on multiple challenging datasets. Furthermore, our method produces these results at around 10Hz, which is many times faster than the related methods. For the sake of reproducibility, we make our trained networks and detection code publicly available.