3D Object Detection and 6D Pose Estimation Using RGB-D Images and Mask R-CNN

Van Luan Tran
Department of Electrical Engineering
National Chung Cheng University
Chiayi 621, Taiwan
tranvanluan07118@gmail.com

Huei-Yung Lin
Department of Electrical Engineering
National Chung Cheng University
Chiayi 621, Taiwan
lin@ee.ccu.edu.tw

***Abstract*—Understanding 3D scenes have attracted significant interests in recent years. Specifically, it is used with visual sensors to provide the information for a robotic manipulator to interact with the target object. Thus, 6D pose estimation and object recognition from point clouds or RGB-D images are important tasks for visual servoing. In this paper, we propose a learning based approach to perform 6D pose estimation for robotic manipulation using Mask R-CNN and the structured light technique. The proposed technique optimizes the 6D pose between the target objects and 3D CAD models in multi-layers. Our method is evaluated on a publicly available dataset for 6D pose estimation and shows its efficiency in computation time. The experimental results demonstrate the feasibility of the random bin picking application.**

***Index Terms*—3D pose estimation, RGB-D image, structured light system, object detection**

I. Introduction

In recent years, deep learning has shown its effectiveness for robot vision, especially with object detection and scene understanding. On the other hand, with the rapid progress in warehouse automation technologies, 6D pose estimation is an important task for a robotic manipulator to determine the exact position of the target object [10]. For instance, recognizing the 3D position and orientation of the objects in a scene is essential for a robotic manipulator. It is also useful in robot vision interaction tasks such as learning from real environments. Nevertheless, the problem is challenging due to the changing position of objects in the real environment. They appear to have different 3D shapes due to the lighting conditions and distortion, and their appearances on RGB-D images are affected by the occlusion between different objects.

Traditionally, 6D pose estimation of an object is performed by matching feature points between the 3D model and the 3D scene [4]. However, these methods require that there are plenty of features on the objects to be detected for matching. As a result, they have the limitation for feature-poor or textureless objects. Currently, with the development of deep learning, the state-of-the-art 3D recognition systems usually adopt two strategies: (1) Recognize the object in a scene with RGB images and project the 3D CAD model of the object to determine its 6D pose [14], [23]. (2) Use 3D deep learning methods for detection, segmentation, and classification to understand the 3D scene with RGB-D images and point clouds [15], [20].

Recently, 3D deep learning has become very popular for object detection and pose estimation by understanding the 3D scene. Specifically, Zeng *et al.* proposed a multi-view self-supervised deep learning technique for 6D pose estimation in the Amazon Picking Challenge [25]. They segmented and labeled multiple views of a scene with a fully convolutional neural network, and then fitted pre-scanned 3D object models to the resulting segmentation to derive the 6D object pose. Training a deep neural network for segmentation typically requires a large amount of training data. Typical computer vision algorithms operate on single images for segmentation and recognition. The robotic arms free us from that constraint by allowing us to precisely fuse multiple views to improve the performance in cluttered environments. However, deep learning in 6D object pose estimation approaches requires significant improvement for the development to make it simple and accurate.

In this paper, we propose an efficient technique to perform 6D pose estimation using the framework summarized in Fig. 1. It is based on Mask-RCNN with the input of RGB images and depth information. Mask-RCNN is a deep neural network trained to solve instance segmentation problem in computer vision [6]. The network detects objects and separates different objects in an image while concurrently producing a high-quality segmentation mask for each object. With the mask of each object, we accumulate the objects in multiple layers and then apply 3D triangulation from a structured light system to calculate 3D point cloud of each object. For the 6D pose estimation in multi-layer, we use a 3D CAD model to optimize the object pose with the ICP algorithm. It is used to find the best transformation that minimizes the distance from the multi-layer source point clouds with the 3D CAD model.

II. Related Work

The main research topics of robot vision include accurate indoor object positioning systems for robotic manipulators (positioning the object), sensor based safety systems, the interaction between human and robot (machine vision), higher levels of realism in vision robot system (3D segmentation,

978-1-7281-6932-3/20/$31.00 ©2020 IEEE

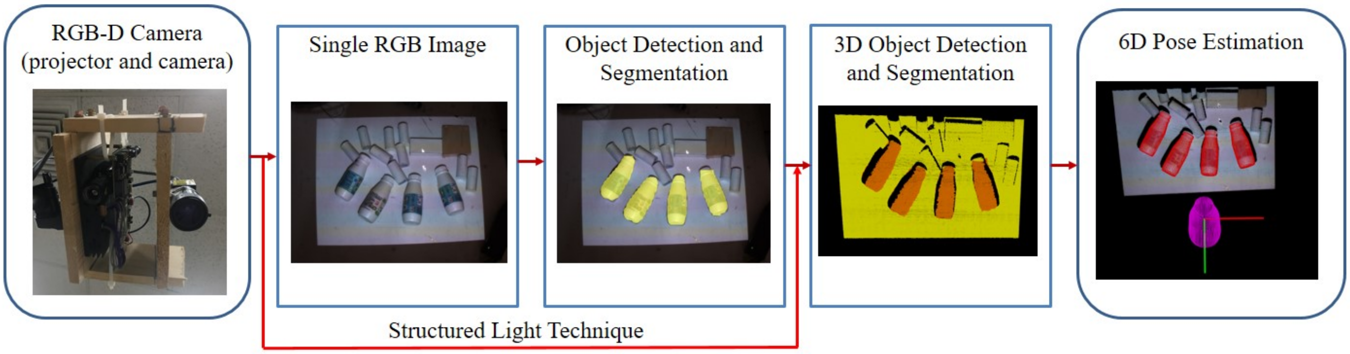

Fig. 1. The overview of our 6D pose estimation based on Mask R-CNN and the structured light technique. The RGB-D camera we developed using a gray-code pattern. From a single RGB-D image, object detection and segmentation are performed with Mask R-CNN. The 6D pose estimation is a result of the estimation between the objects in multiple layers and the 3D CAD models.

classification, recognition, and 6D pose estimation), and reactive planning and controllability in real industrial factory or workshop safety (machine learning) [11]. In this work, we investigate the existing methods related to object recognition and 6D pose estimation for robot vision using either traditional or deep learning methods.

In the existing literature, a number of 6D object pose estimation and 3D recognition techniques have been proposed. Recent state-of-the-art 6D object pose estimation systems based on deep learning usually adopt two main strategies: (1) The approaches using RGB and depth images for training and testing, to detect, segment and render the object with a depth image to optimize the 6D pose [8], [21]; (2) The approaches using XYZRGB point clouds for training and testing, to perform 3D classification, segmentation and optimize the 6D pose, or directly estimate the pose using 3D convolution on the point clouds [2], [5].

A. Deep Learning on Images

In [24], Xiang *et al.* propose a new convolutional neural network for 6D object pose estimation (PoseCNN). It learns to predict object labels for semantic labeling, and the object location and orientation. To improve the 6D object pose estimation, a modified ICP algorithm is carried out on the depth images using the output of PoseCNN (PoseCNN+ICP). PoseCNN is low inaccuracy due to the lack of depth information, but it can be performed in short computation time. When PoseCNN is combined with a modified ICP algorithm, the accuracy increases at a cost of large computation time. Pavlakos *et al.* present a novel approach to estimate the continuous 6D pose of an object from a single RGB image [14]. They combine semantic keypoints predicted by a convolutional network (convnet) with a deformable shape model. In [7], He *et al.* propose a 3D object detection and pose estimation pipeline using RGB-D images, which can detect multiple objects simultaneously while retaining a low false positive rate. The approach starts with template matching, generating a set of matches as initial results. They achieve an average success rate of 83.41% and limit the feasibility of some grasping poses. For example, in pick-and-place applications, objects in certain poses do not expose enough areas for suction, which might cause major picking failures.

B. Deep Learning on Point Clouds

Currently, neural networks are trained with 3D object datasets, and multi-layer neural networks are able to learn features automatically using a general purpose training model by 3D convolutional neural network. In [16], Qi *et al.* address the task of object classification on 3D data using volumetric CNNs and multi-view CNNs. They analyze the performance gap between volumetric CNNs and multi-view CNNs from the perspectives of network architecture and 3D resolution. Recently, m multi-view convolutional neural networks [18] have been proposed with significant results for 3D shape recognition. By a novel CNN architecture that combines information from multiple views of a 3D shape into a single compact shape descriptor, they offer even better recognition performance. However, there are still some problems with how many views are necessary for a given level of accuracy. Qi *et al.* propose a PointNet, which is an important type of geometric data structure for the point cloud. It is highly efficient for object classification, part segmentation, and scene semantic representation. With 3D shape models, 3DShapeNet (a convolutional deep belief network) [22] learns the joint distribution of generic 3D shapes across object categories. They train a large-scale 3D CAD model dataset with volumetric representation for 3D CNNs to perform 3D object recognition. These significant 3D deep learning methods can achieve an accuracy of 83.54%. Using 3D convolution on XYZRGB point clouds for object pose estimation, Cai *et al.* propose a method to estimate the object's translation and rotation with the translation error around 1 cm and the rotation error around 5 degrees [2]. It achieves an above 90% success rate for test objects using a robotic manipulator.

III. APPROACH

The problem in this work concerns estimating the 6-DOF pose of an object located in a scene, with the inputs given by:

- RGB images and depth information provided by a structured light system,
- a set of 3D CAD models of the target objects, one for each object,
- and the 6-DOF pose of the camera system.

The outputs are the rotation matrix and translation vector, which represent the relationship between the target object and its 3D CAD model. The proposed approach consists of three parts: (1) an RGB-D camera system developed based on the structured light technique, (2) detection, segmentation and the 3D information derivation of target objects, and (3) 6D pose estimation from the correspondence between the objects and CAD models.

A. Structured Light Technique

RGB-D cameras have been widely used to perform the 6D pose estimation of target objects for robotic manipulators. One popular 3D acquisition approach is to adopt a structured light system. The underlying technology is based on the principle described in Fig. 2. In general, the process of a structured light 3D acquisition system can be divided into three basic steps:

- **Encoding:** Encoding the 3D information into a sequence of patterns is performed in the temporal domain.
- **Acquisition:** A sequence of structured light patterns is projected to the 3D scene by a data projector and continuously captured by a camera.
- **Decoding:** The captured pattern-coded images are processed to decode and find the corresponding points associated with the projector and camera.

In the implementation, there might be some additional steps depending on the solution of the system developer. It often follows a procedure to create range images, point clouds or mesh models, possibly integrate several decoded coordinate maps, calibration, and the triangulation principle. The calibration is to determine the intrinsic and extrinsic parameters of the camera and projector, and the reconstruction is usually based on the ray triangulation principle by computing the intersection point.

We have developed a structured light RGB-D camera for accurate depth measurement with gray-code coding [19]. The 3D reconstruction is achieved by the ray triangulation principle with the estimation of intersection points. The accuracy and density of the obtained point clouds are both high, and therefore it is suitable for applications such as accurate shape measurements, 3D object recognition, and pose estimation for robotic manipulation.

B. Object Detection and Segmentation

For the real-time object detection, Ren *et al.* propose a faster R-CNN algorithm [17]. It is a region proposal network (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. An RPN is a fully convolutional network that simultaneously predicts object bounds and objectness scores at each position, which reduced the running time of the detection networks. To solve the instance segmentation problem in computer vision, He *et al.* propose Mask R-CNN (Regional Convolutional Neural Network) [6] with a high-quality segmentation mask for each instance. It efficiently detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance. The network extends Faster R-CNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition.

Mask R-CNN is a two-stage structure: the first stage scans the input image about the regions and generates proposals where there might be an object. The second stage predicts the class of the object, refines the bounding box, and makes a mask in the pixel level of the object based on the first stage proposal. The output of this stage is a binary mask for each ROI align in parallel to predict the class. The ROI is to determine the relevant areas of the feature map, and there is a branch generating masks for each object in pixel level. Particularly, the mask is quickly picked from the label of each ROI classification in parallel by the dedicated classification category. Both stages are connected to the backbone structure. The backbone is a Feature Pyramid Networks (FPN) style deep neural network. FPN applies a top-down architecture with lateral connections to establish an in-network feature pyramid from a single-scale input. FPN outperforms other single ConvNets mainly for the reason that it maintains strong semantically features at various resolution scales. The structure of the Mask R-CNN is described in Figs. 3 and 4. The object classification is first segmented on each layer and obtain the 3D information for the object by the structured light system. The number of the layer depends on the number of the object for classification.

C. 6D Pose Estimation

The object detection and segmentation on each layer will obtain the 3D information and directly perform the 6D pose estimation. To estimate the 6D object pose, we use the ICP algorithm. It is a robust algorithm for calculating the transformation between the object in the scene and the model. The algorithm aligns the 3D models and targets based on geometry, and is widely used to register the outputs of 3D scanners and 3D reconstruction for 6D pose estimation. ICP originates with two point clouds and an initial guess for their relative rigid body transformation. The major problem is to determine the correct data associations. Given the correct data associations, the transformation can be computed efficiently using a singular value decomposition (SVD).

Given two corresponding point sets of the model and the object by

$$M = \{m_1, ..., m_n\}, \quad P = \{p_1, ..., p_n\} \tag{1}$$

The translation vector t and rotation matrix R that minimize the sum of the squared error:

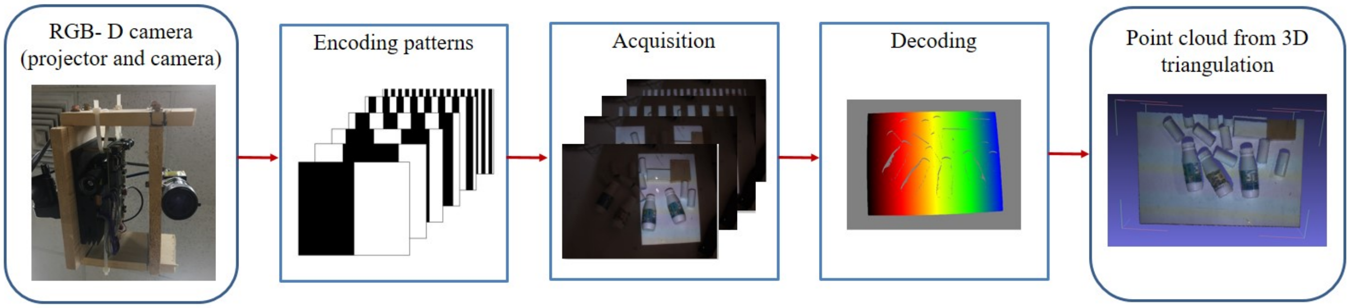

Fig. 2. The RGB-D camera system with the structured light technique. Our RGB-D camera uses one projector and one camera. The encoding pattern is a gray-code pattern. The acquisition is the images captured by the camera for each pattern in the sequence. The decoding is a coded map. The 3D point cloud is derived based on the ray triangulation principle by the estimation of intersection points

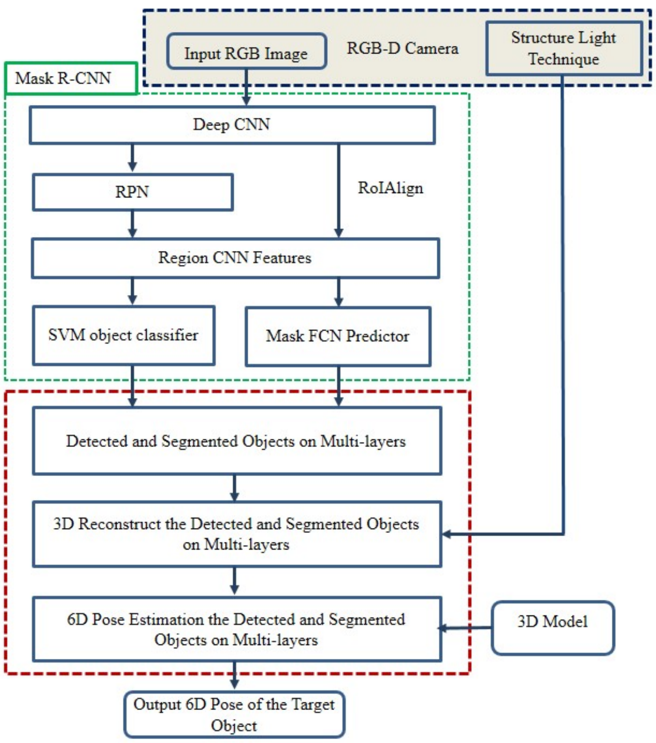

Fig. 3. The structure of our solution to perform 6D pose estimation based on Mask R-CNN and the structured light technique. With Mask R-CNN structure, the detected and segmented objects are accumulated in multiple layers. The 6D pose estimation of the detected and segmented objects is performed between a single layer and 3D CAD models for all object layers with the ICP algorithm.

$$E(R,t) = \frac{1}{N_p}\sum_{i=1}^{N_p} ||m_i - Rp_i - t||^2 \tag{2}$$

where m_1 and p_1 are the corresponding points.

In our system, the single 3D object is stored in a single layer by 3D object detection and segmentation. Then, ICP uses the single 3D object in each layer to match the 3D CAD model. The centers of mass of the two-point sets (model and object) are:

$$C_m = \frac{1}{N_m}\sum_{i=1}^{N_m} m_i, \quad C_p = \frac{1}{N_p}\sum_{i=1}^{N_p} p_i \tag{3}$$

We subtract the corresponding center of mass from every point in the two-point sets before calculating the transformation. The resulting point sets are:

$$M' = \{m_1 - C_m\} = \{m'_1\} \tag{4}$$

$$P' = \{p_1 - C_p\} = \{p'_1\} \tag{5}$$

Let

$$A = \sum_{i=1}^{N_p} m'_i p'^T_i \tag{6}$$

denote the SVD of A by

$$A = U \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{bmatrix} V^T \tag{7}$$

where $U, V \in R^{3\times3}$ are unitary, and $\sigma_1 \geq \sigma_2 \geq \sigma_3$ are the singular values of A. The optimal solution of $E(R,t)$ is unique and is given by

$$R = UV^T, \quad t = C_m - RC_p \tag{8}$$

IV. EXPERIMENTS

Our structured light RGB-D camera system consists of a Flea3 FL3-U3-32S2C camera from Point Grey Research with the image resolution of 2080×1552. The digital light projector is a DLP Light Crafter 4500 projector from Texas Instrument with a resolution of 1024×678. The camera is installed at about $1m$ from the table and with a 30-degree angle looking down.

We apply the pre-trained network [1] to our own dataset with the Mask R-CNN model [6]. The objects detected and segmented by Mask R-CNN are divided into multiple layers. The number of layers depends on the number of the detected object as presented in Fig. 4. In this experiment, we place four bottles on a table and they are detected and segmented to

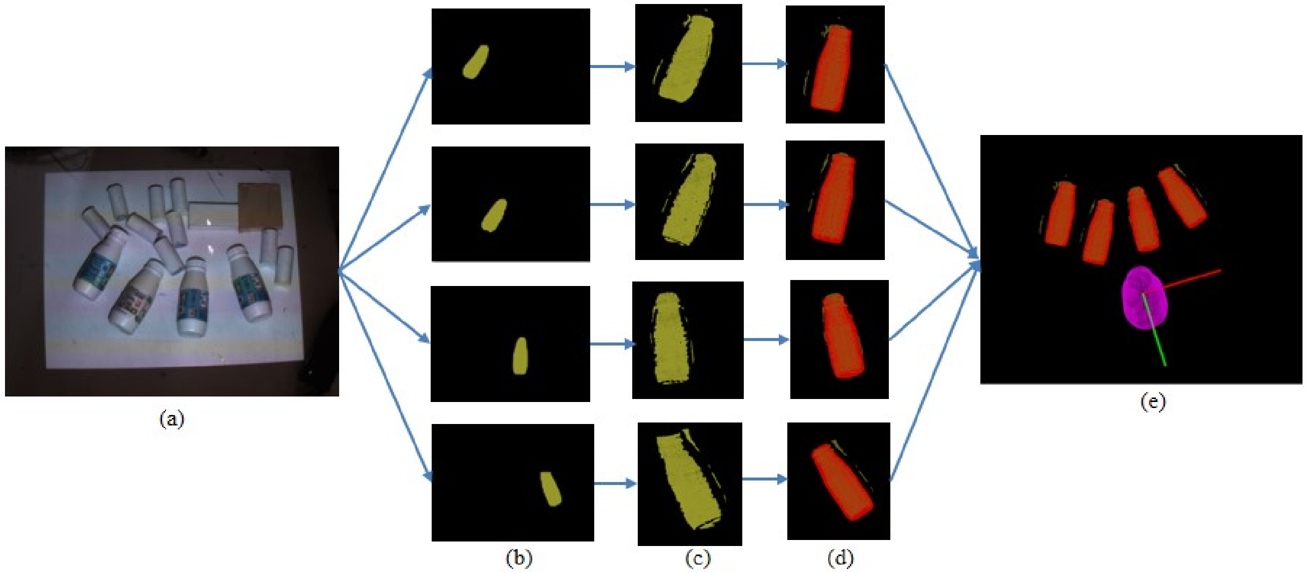

Fig. 4. The overview of our approach for 6D pose estimation: (a) one input RGB image from the RGB-D camera, (b) object detection and segmentation in multi-layer based on Mask R-CNN, (c) 3D detected objects in the multiple layers with the 3D information from the structured light system. (d) estimating the 6D object pose by the ICP algorithm in multi-layer, and (e) output results with the object pose and the 3D CAD model.

four layers. In each layer, the 3D information obtained from the structured light system is directly used for pose estimation with 3D CAD models. Each layer is processed in a parallel way to reduce the cost of the computation.

In the experiment, we set up the scene with the objects placed randomly in front of the camera with different angles, as shown in Fig. 5. The target objects are recognized and the rotation matrix and translation vector are obtained for each one. In the figure, we show the 6D object pose in red and the 3D CAD model arranged at the RGB-D camera in purple. The 6D object pose estimation is evaluated as the percentage of prediction with a translation error less than 1 cm and the mean error in rotation less than 3 degrees.

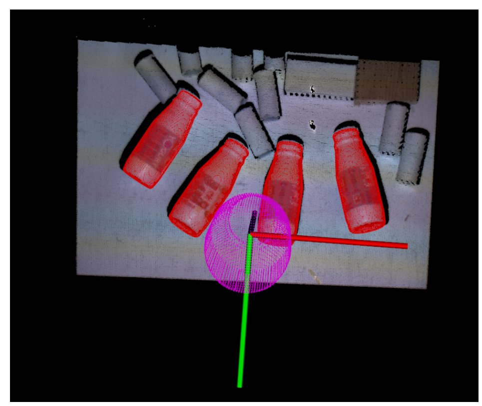

Fig. 5. The result of 6D object pose estimation with four bottles placed on a table with some other objects. The four red objects are the output of 6D pose presented in a full 3D scene.

In Fig. 6, we test our method using some bottles placed in a box mixed with other objects to make the scene more complicated. The object with unsuccessful pose estimation will be removed by the fitness score of the ICP algorithm. In the experimental result, we have selected two objects with the best pose estimation.

We have evaluated our method with the YCB-video dataset [3], a standard dataset for robotic manipulation tasks. It presents accurate 6D poses of 21 objects captured from multiple views and annotation of 6-DOF object poses with 133,827 frames. As shown in Table I, our proposed method achieves 87.4% accuracy. It outperforms with other methods [13], [9], except PoseCNN+ICP. However, in our system, the computation time can be reduced at an average runtime of 1.4 seconds per object. with PoseCNN+ICP, the accuracy increases at a cost of large computation time (approximately 10.6 seconds per object for the modified-ICP refinement [12]). The main reason is that we only find the 3D information of a single object with semantic labeling and directly estimate the 6D object pose by an ICP algorithm in a multi-layer.

V. CONCLUSION

In this paper, we propose a structural hybrid technique for 3D recognition based on Mask R-CNN and structured light 3D acquisition. With the object detection and segmentation based on Mask R-CNN and accurate 3D information from the structured light camera system, our approach is able to make the 6D pose estimation more robust. We use the 3D CAD model to optimize the object pose with the ICP algorithm and achieve low computation time with the 6D pose estimation in a multi-layer. The experiments using the real-world images have been carried out and the performance is evaluated and compared with several state-of-the-art methods. The results have demonstrated the effectiveness of our technique.

TABLE I

SOME RESULTS FOR THE POSE ESTIMATION ON YCB-VIDEO DATASET [3].

Method	Input	Supervision	Real-images	Accuracy (%)
PoseCNN+ICP [24]	RGBD	6D pose labels	113,199	93.0
PoseCNN [24]	RGB	6D pose labels	113,199	75.9
DeepHeatmaps [13]	RGB	6D pose labels	113,199	66.1
Mercier et. al. [9]	RGBD	Object class labels	210	83.9
OURS	RGBD	6D pose labels	113,199	87.4

Fig. 6. The result of 6D object pose estimation with some bottles placed in a box with other objects, The two red objects are the output of 6D pose presented in a full 3D scene.

ACKNOWLEDGMENT

The support of this work in part by the Ministry of Science and Technology of Taiwan under Grant MOST 106-2221-E-194-004 and the Advanced Institute of Manufacturing with High-tech Innovations from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education in Taiwan is gratefully acknowledged.

REFERENCES

[1] W. Abdulla. Mask r-cnn for object detection and instance segmentation on keras and tensorflow, 2017.

[2] Z. Cai, C. Yu, and Q. Pham. 3d convolution on RGB-D point clouds for accurate model-free object pose estimation. *CoRR*, abs/1812.11284, 2018.

[3] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 International Conference on Advanced Robotics (ICAR)*, pages 510–517, July 2015.

[4] K. Cheng, H. Lin, and T. V. Luan. A 3d recognition system with local-global collaboration. In *Computer Vision - ACCV 2016 Workshops - ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II*, pages 19–33, 2016.

[5] F. Engelmann, T. Kontogianni, J. Schult, and B. Leibe. Know what your neighbors do: 3d semantic segmentation of point clouds. *CoRR*, abs/1810.01151, 2018.

[6] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[7] R. He, J. Rojas, and Y. Guan. A 3d object detection and pose estimation pipeline using RGB-D images. In *2017 IEEE International Conference on Robotics and Biomimetics, ROBIO 2017, Macau, Macao, December 5-8, 2017*, pages 1527–1532, 2017.

[8] O. H. Jafari, S. K. Mustikovela, K. Pertsch, E. Brachmann, and C. Rother. The best of both worlds: Learning geometry-based 6d object pose estimation. *CoRR*, abs/1712.01924, 2017.

[9] J. Mercier, C. Mitash, P. Giguère, and A. Boularias. Learning object localization and 6d pose estimation from simulation and weakly labeled real images. *CoRR*, abs/1806.06888, 2018.

[10] C. Mitash, K. E. Bekris, and A. Boularias. A self-supervised learning system for object detection using physics simulation and multi-view pose estimation. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 545–551, Sep. 2017.

[11] C. Mitash, A. Boularias, and K. E. Bekris. Improving 6d pose estimation of objects in clutter via physics-aware monte carlo tree search. In *2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018*, pages 1–8, 2018.

[12] C. Mitash, A. Boularias, and K. E. Bekris. Robust 6d object pose estimation with stochastic congruent sets. *CoRR*, abs/1805.06324, 2018.

[13] M. Oberweger, M. Rad, and V. Lepetit. Making deep heatmaps robust to partial occlusions for 3d object pose estimation. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*, pages 125–141, 2018.

[14] G. Pavlakos, X. Zhou, A. Chan, K. G. Derpanis, and K. Daniilidis. 6-dof object pose from semantic keypoints. In *2017 IEEE International Conference on Robotics and Automation, ICRA 2017, Singapore, Singapore, May 29 - June 3, 2017*, pages 2011–2018, 2017.

[15] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CoRR*, abs/1612.00593, 2016.

[16] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. J. Guibas. Volumetric and multi-view cnns for object classification on 3d data. *CoRR*, abs/1604.03265, 2016.

[17] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, June 2017.

[18] H. Su, S. Maji, E. Kalogerakis, and E. G. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 945–953, 2015.

[19] V. Tran and H. Lin. Accurate rgb-d camera based on structured light techniques. In *2017 International Conference on System Science and Engineering (ICSSE)*, pages 235–238, July 2017.

[20] P. Wohlhart and V. Lepetit. Learning descriptors for object recognition and 3d pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3109–3118, 2015.

[21] J. M. Wong, V. Kee, T. Le, S. Wagner, G. L. Mariottini, A. Schneider, L. Hamilton, R. Chipalkatty, M. Hebert, D. M. S. Johnson, J. Wu, B. Zhou, and A. Torralba. Segicp: Integrated deep semantic segmentation and pose estimation. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2017, Vancouver, BC, Canada, September 24-28, 2017*, pages 5784–5789, 2017.

[22] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1912–1920, 2015.

[23] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *CoRR*, abs/1711.00199, 2017.

[24] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Robotics: Science and Systems XIV, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 26-30, 2018*, 2018.

[25] A. Zeng, K. Yu, S. Song, D. Suo, E. W. Jr., A. Rodriguez, and J. Xiao. Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge. In *2017 IEEE International Conference on Robotics and Automation, ICRA 2017, Singapore, Singapore, May 29 - June 3, 2017*, pages 1386–1383, 2017.