View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Cronfa at Swansea University

Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 23, 2020, accepted March 16, 2020, date of publication March 25, 2020, date of current version April 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2983173

Fast Object Pose Estimation Using Adaptive Threshold for Bin-Picking

WU YAN[1], ZHIHAO XU[1,3], (Student Member, IEEE), XUEFENG ZHOU[1], QIANXING SU[1], SHUAI LI[2], (Member, IEEE), AND HONGMIN WU[1]

[1]Guangdong Key Laboratory of Modern Control Technology, Guangdong Institute of Intelligent Manufacturing, Guangzhou 510070, China
[2]School of Engineering, Swansea University, Swansea SA2 8PP, U.K.
[3]Foshan Tri-Co Intelligent Robot Technology Company, Ltd., Foshan 528315, China

Corresponding author: Hongmin Wu (hm.wu@giim.ac.cn)

This work was supported in part by the Foshan Key Technology Research Project under Grant 1920001001148, in part by the GDAS' Project of Thousand doctors (post-doctors) Introduction under Grant 2020GDASYL-20200103128, in part by the Foshan Innovation and Entrepreneurship Team Project under Grant 2018IT100173, in part by the Guangdong Province Key Areas Research and Development Program under Grant 2019B090919002, in part by the Guangzhou Science Research Plan—Major Project under Grant 201804020095, in part by the Guangdong Innovative Talent Project of Young College under Grant 2016TQ03X463, and in part by the Guangzhou Science and Technology Plan Project under Grant 201803010106.

ABSTRACT Robotic bin-picking is a common process in modern manufacturing, logistics, and warehousing that aims to pick-up known or unknown objects with random poses out of a bin by using a robot-camera system. Rapid and accurate object pose estimation pipelines have become an escalating issue for robot picking in recent years. In this paper, a fast 6-DoF (degrees of freedom) pose estimation pipeline for random bin-picking is proposed in which the pipeline is capable of recognizing different types of objects in various cluttered scenarios and uses an adaptive threshold segment strategy to accelerate estimation and matching for the robot picking task. Particularly, our proposed method can be effectively trained with fewer samples by introducing the geometric properties of objects such as contour, normal distribution, and curvature. An experimental setup is designed with a Kinova 6-Dof robot and an Ensenso industrial 3D camera for evaluating our proposed methods with respect to four different objects. The results indicate that our proposed method achieves a 91.25% average success rate and a 0.265s average estimation time, which sufficiently demonstrates that our approach provides competitive results for fast objects pose estimation and can be applied to robotic random bin-picking tasks.

INDEX TERMS Robotic bin-picking, learning-based pose estimation, adaptive threshold, point pair features, multilayer segmentation, 3D sensing.

I. INTRODUCTION

An accurate, fast, and robust 6D object pose estimation solution has served an important role in many practical applications such as robot manipulation, augmented reality and autonomous driving. The RGB-D camera (aka 3D sensor or depth camera) is a critical device for obtaining visual information from commercial to industrial levels, such as Kinect, RealSense, Bumblebee, etc., which has been proven in deriving better performances in 6D object pose estimation with respect to the RGB cameras without depth or distance measurement. Because the depth cameras concurrently capture both the appearance and geometry of the scene, those RGB-D sensors also have capabilities for accurately inferring poses of low-textured objects, even in poorly illuminated environments. Additionally, an effective method for building a database of multiple objects in random poses is desirable, which improves the robustness of pose estimation for random bin-picking in industrial applications, as shown in Figure 1. The former is textureless, and the later is thin and partially occluded. However, this consideration still presents challenges to be addressed, including robustness against occlusion and clutter, scalability to multiple objects as well as the fast and reliable object modeling methods. To address this problem, two categories of mainstream methods were presented for fast objects pose estimation.

A. FEATURE-BASED METHODS

Conventional methods commonly use key-point descriptors or some heuristic features (point, line, edge, etc.) of those captured images, where the descriptor includes transformation

The associate editor coordinating the review of this manuscript and approving it for publication was Xiwang Dong.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

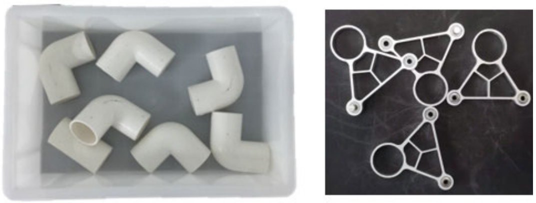

FIGURE 1. Objects are placed for robot grasping in industrial applications.

from 2D descriptors (SIFT [1], SURF [2], ORB [3], etc.) to 3D local and global feature descriptors (FPFH [4], [5], USC [6], SHOT [6], [7], Spin Image [8], CVFH [9], GRSD [10], [11], etc.). Those methods were designed according to a common pipeline by performing correspondence grouping, hypothesis verification, ICP [12]–[14] as well as the refinement on the last pose for obtaining a more accurate result [15]–[20]. However, the 6D pose estimation methods using 2D/3D images in robotic random bin-picking systems only work when the poses are limited to a few degrees of freedom and suffer from significant lighting changes or textureless objects. Particularly, Ulrich *et al.* [16] proposed a method that uses 2D edge-based matching and a pyramid-level fashion to iteratively search for the maximum similarity in a 2D model in order to estimate the 3D pose of an object. In [17], the author first attempted to use the 2D key-point matching method for implementing the robotic manipulation of 3D objects resulting in a new solution for pose estimation. As for 3D point cloud methods, Song *et al.* [21] proposed a multiview pose estimation system, but the methods to create the database are time-consuming and not very fast for bin-picking tasks.

Two methods that must be mentioned are the multimodal-LINE (LINEMOD) algorithm and point pair feature (PPF) algorithm proposed by Hinterstoisser *et al.* [22] and Drost *et al.* [23], respectively. LINEMOD achieves a high recognition rate and speed on the ACCV3D dataset by combining color gradient and normals information. The PPF algorithm only uses depth information, combines an efficient voting scheme with point pair features and does not use color information: both of them are robust to partial occlusion. PPF has exhibited good pose estimation performance on multiple datasets, winning the SIXD challenge in 2017 [24], and many modifications have been made to the original PPF [25]–[31]. Hinterstoisser *et al.* [25] introduced a better and more efficient sampling strategy with modifications to the pre and post-processing steps, which achieved good results. Liu *et al.* obtained an impressive result by combining a machine learning-based 2D object localization and a 3D pose estimation method in [26], [27]. Vidal *et al.* proposed a better preprocessing and clustering step with an improved matching method for Point Pair Features in [29]. Li *et al.* proposed a novel descriptor Curve Set Feature (CSF) and a method named Rotation Match Feature (RMF) to speed up the matching phase in [30] and used foreground and background segmentation with synthetic scenes to make up for the point pair feature algorithm in [31].

B. LEARNING-BASED POSE ESTIMATION

Convolutional Neural Networks (CNNs) have revolutionized object detection from RGB images. More recent deep learning architectures have performed detection on 3D data directly [32]–[35]. PointNet [34], [35] is the pioneer which fed raw point cloud data into neural networks. PoseCNN [36] is trained to predict 6D object poses from images in multiple stages, by decoupling the translation and rotation predictors. BB8 [37] and SSD-6D [38] do not directly predict object pose; instead, they perform all predictions for 2D images. DeepIM [39] is a novel framework for iterative pose matching using color images only. DenseFusion [40] introduces a neural network that combines RGB images and depth images for 6D pose estimation, with an iterative pose refinement network using point clouds as input. DeepHMap++ [41] introduces a novel method for robust and accurate 3D object pose estimation from single-color images with significant occlusion. Nevertheless, the majority of learning-based pose estimation methods use real labeled images and are thus restricted to pose-annotated datasets [37], [42]. Consequently, several related works [38], [43], [44] have been proposed to train on synthetic images rendered from a 3D model, yielding a great data source with pose labels free of charge. More recently, automatic annotation methods have been proposed using motion capture [45] or 3D scene reconstruction [46], [47]. However, in comparison to 2D bounding box annotation, the effort of labeling real images with full 6D object poses is more difficult and requires expert knowledge with a complex setup [48]. These methods still require significant human labor and are not able to generate variability in pose since objects must remain stationary during data capture.

However, naive training on synthetic data does not typically generalize to real test images. Therefore, the main challenge is to bridge the domain gap that separates simulated views from real camera recordings. In this paper, we propose a fast CAD-based 6-DoF pose estimation pipeline for random bin-picking for different objects of varying shapes. 3D CAD data or target partial point clouds were used to create an off-line database that includes different feature information (point, surface normal) about the model. Concerning the pose estimation task, a voting-based feature matching scheme was introduced to perform comparison scheduling using the target in the database and real scenes to create pose clustering. Last, but equally importantly, RANSAC (random sample consensus) and ICP (iterative closest point) are used to pose refinement to obtain a more accurate result. Our main contribution in this paper is the fast multilayer picking strategy in the bin-picking problem and the method to create a new offline database with fewer samples and less training time when unseen types of objects are added.

The remainder of this paper is organized as follows. First, the system architecture for CAD-based pose estimation is described in Section II. We subsequently explain details of the feature extractor, voting-based matching procedure and multilayer adaptive threshold picking strategy in Section III. With the proposed pipeline, a real-world experiment is

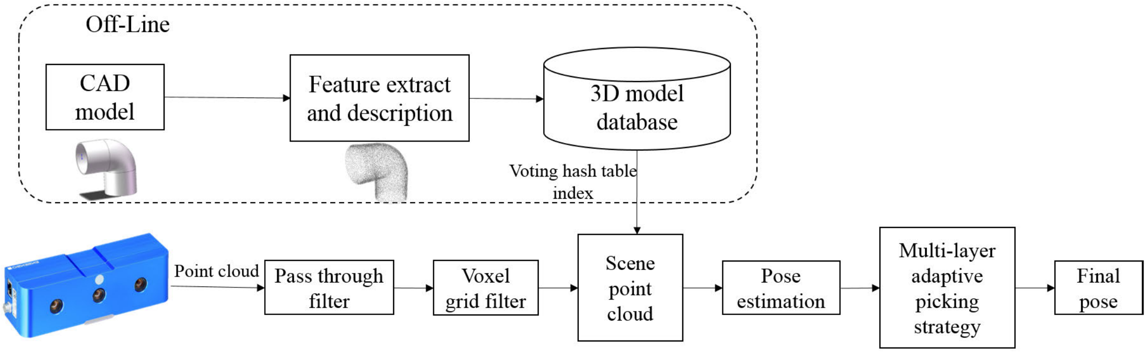

FIGURE 2. **Illustration of a proposed pipeline for 3D object detection and pose estimation. The 3D model database in the dotted rectangle is built off-line, followed by feature extraction and point cloud key point description transform of the 3D CAD model into the 3D model database as template.**

presented in Section IV for evaluating the proposed methods, and the experimental results and discussions are also included for supporting the performance of the proposed random bin-picking system. Finally, we present the conclusion and future direction in Section V.

II. SYSTEM ARCHITECTURE

The proposed system architecture for robot bin-picking is shown in Figure 2, which consists of three main components:

- The establishment of the off-line 3D model database;
- Voting-based matching procedure and RANSAC 6-DoF pose estimation;
- Target selection using multilayer adaptive threshold.

As illustrated in Figure 2, those operations surrounded by dotted lines are in the offline phase primarily for creating a 3D model database by extracting features from the 3D CAD model or point clouds from real scenes, in which the models' point features can easily be obtained and automatically generate informative samples by randomly adjusting the customized parameters (scale, position, and orientation). In addition, it can not only use CAD data but also the real point cloud, as shown in Figure 3, respectively. The left figure illustrates the 3D CAD model used to generate offline datasets, where the yellow dots are key points and the reference points are in green within a certain distance between the key points, the template model would consist of contour, normal distribution, and curvature information. The right figure shows the datasets generated by using the scene segmentation method in real-time point cloud when the CAD model is hard or time-consuming to get. The yellow and green dots are the same as above. The feature generation will be discussed in detail in Section III.

The 3D camera captures depth images and generates point clouds for the scene. To speed up the processing time, pass-through filter and voxel grid filter was used to remove noise, the point cloud then followed with a voting-based matching scheme which aims to determine the correspondence between the offline database and the captured scene. The pose estimation system estimated different types of objects with their 6-DoF poses in the bin. After that, the multilayer adaptive picking strategy was introduced to obtain the best target and its' pose which is most suitable to allow the robot gripper to pick it up among all of the recognized candidate objects.

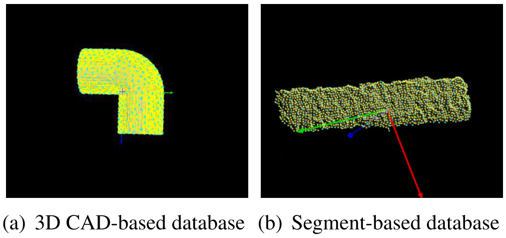
(a) 3D CAD-based database (b) Segment-based database

FIGURE 3. **Dataset construction (a) An instance of the dataset is created from the 3D CAD model, where the key points and the reference point are in yellow and green, respectively. (b) By using the segmentation method in real-time point cloud, another instance (computer-memory-card) in dataset is created, the color scheme is the same as previous.**

III. POSE ESTIMATION AND ADAPTIVE THRESHOLD SEGMENTATION

A. OBJECT DATASET CONSTRUCTION

A fast and user-friendly pipeline to construct object dataset is described in this section. Usually, it's not difficult to get the corresponding 3D model because the workpiece has its own 3D model before it is made. At the same time, compared to setting virtual camera in space and taking snapshots from different viewpoints [16], [49], our methods does not require setting multiple viewing angles, includes no blind zone and will automatically generate informative samples by randomly adjusting the customized parameters (scale, position, and orientation), which can perfectly obtain the complete information of the model and contain more information. To build a description of the 3D CAD model, the point cloud data are converted to the Point Pair Features (PPF) [23] descriptors and then similar features are grouped in a hash table. The hash table is a representation of the 3D CAD model that is stored in the 3D model database. The 3D model database is then used to estimate the 6-DoF pose, in order to recognize and localize objects.

In addition, as for some model has difficulty in obtaining the 3D CAD data, we have proposed a fast and user-friendly way to learn it online such that there are two ways

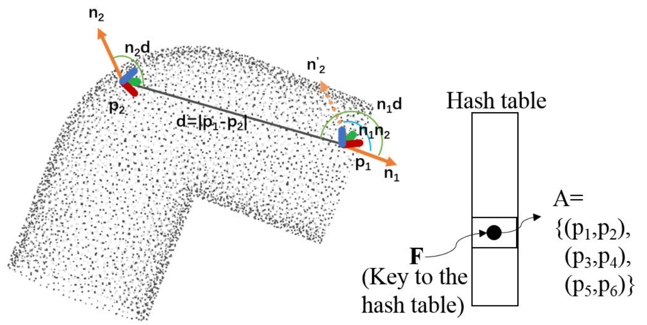

FIGURE 4. **Illustration of the *PPF* descriptor on the point cloud.**

to create an off-line database in our system, that is the major difference compared with others. Depth-based methods of PPF have shown robust pose estimation performance on multiple datasets. Whereas, we use the voting-based PPF algorithm [50] to perform point sampling and pose voting. The PPF descriptor is a 4-dimensional descriptor, formulated by Equation (1), which encodes the relations between two surface points, $p1$ and $p2$, with its normal vector.

$$\mathbf{PPF}(\mathbf{p_1}, \mathbf{p_2}) = [\| d \|_2, \angle(n_1, d), \angle(n_2, d), \angle(n_1, n_2)] \quad (1)$$

where d is the Euclidean distance between $p1$ and $p2$, $n1$ and $n2$ are the point normal and $\angle(n_1, n_2)$ are the angles between the point normal, and $\angle(n_1, d)$ and $\angle(n_2, d)$ are direction vectors between the two points, as denoted in Figure 4.

B. VOTING-BASED MATCHING PROCEDURE AND RANSAC POSE ESTIMATION IN MULTIPLE SCENES

The proposed 6-DoF pose estimation system recognizes and localizes different types of workpieces. The system architecture for the pose estimation system is shown in Figure 5. Outside the dotted rectangle, the depth image is the input and a set of 6-DoF poses for different types of objects is the output. The algorithm has an offline phase and an online phase. During the offline phase, the 3D model database is generated for recognition. During the online phase, the implementation procedure of the pose estimation module is as follows:

- Converting the depth image of the scene into the point cloud.
- Outliers in the scene were removed using the pass-through filter to increase processing speed and produce a smaller search area.
- Voxel grid filter was used for downsampling and to decrease the number of point clouds.
- The point-pair feature descriptor was introduced to compute the key-point features and then match them with the 3D model database to estimate the 6-DoF pose using the voting-scheme matching algorithm.
- Using the pose clustering picking algorithm to remove the incorrect 6-DoF pose proposals that have high scores. Since an object is modeled by a large set of pair features, it is expected to have multiple pose hypotheses each for different reference points supporting the same pose. The retrieved poses are clustered such that all poses in one cluster do not differ in translation and rotation by more than a predefined threshold.
- Increasing the accuracy of the 6-DoF pose estimation results by using the ICP algorithm [12]–[14] which produced more accurate results than simulation for 6-DoF pose estimation.

Here, we denote M and S as the CAD model and scene point cloud separately, and $PPF_\Theta(p1, p2)$ and $PPF_\Omega(p1, p2)$ as the Point Pair Feature on the model and scene point cloud Point Pair Feature, as formulated in Equation 2. First, we calculate all possible $PPFs$ within the minimum distance of the model. In addition, we also eliminate the noise points where the distance between them is greater than the predefined value. We can think of this as a training phase. We then use the four parameters to obtain similar point pairs on the model surface, and those point pairs are stored in the same slot in the hash table. We then create an array of votes for every CAD model point pair and vote for every possible correspondence. Finally, we convert the reference point and key point to the same local frame and find the closest $PPFs$ from the object model using a k-d tree with a radius search for each reference point in the scene of Figure 6, the maximum value in the voting space is the locally optimal pose, which could be treated as a matching phase. Additionally, the CAD models are matched by those from the scene feature descriptors,

$$\begin{aligned}\mathbf{M} &= \{PPF_\Theta(p_1, p_2), PPF_\Theta(p_3, p_4), \ldots, PPF_\Theta(p_n, p_m)\}\\ \mathbf{S} &= \{PPF_\Omega(p_1, p_2), PPF_\Omega(p_3, p_4), \ldots, PPF_\Omega(p_n, p_m)\} \quad (2)\end{aligned}$$

The three-dimensional matching is mainly calculated based on the three-dimensional point cloud and the normal. The basic principle is that the matched key points have similar distances and normal information. A point pair $(\mathbf{m}_r, \mathbf{m}_i) \in M^2$ is aligned to a scene pair $(\mathbf{s}_r, \mathbf{s}_i) \in S^2$ where both pairs have a similar feature vector F. The more similar these are, the higher the voting score between them. Then, because the model includes pose information in the data description stage, information among different pose clustering methods can be used to acquire the target pose (e.g., nearest neighbor, RANSAC). In this work, we use a RANSAC algorithm to remove erroneous correspondences and acquire the pose of the target. After that, the pose is refined with the ICP algorithm to produce a more accurate result.

C. MULTILAYER ADAPTIVE THRESHOLD PICKING STRATEGY

For the bin-picking problem in industrial scenarios, the major improvement of production efficiency always is determined by the speed of object recognition and picking. To address this problem, a multilayer adaptive threshold picking strategy is proposed in this paper. The strategy can be easily adopted to the varying shapes of objects because each object has a maximum convex enclosure and that every layer of the gap between enclosures is always less than the diameter of the convex enclosure even in situations with random overlap. Here, prior knowledge is taken into consideration in the robot bin-picking system, and the depth of interest between the

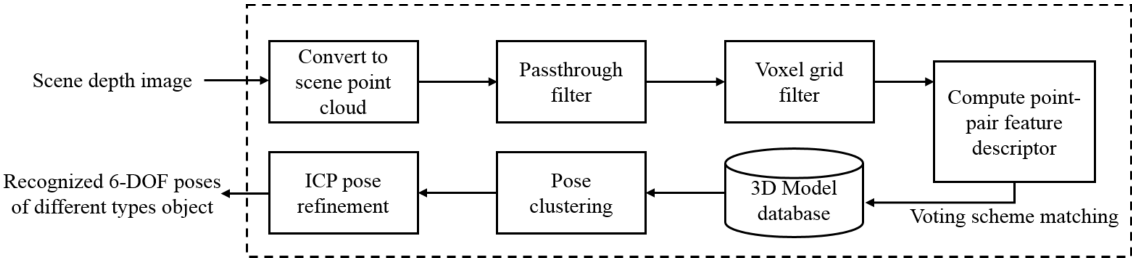

FIGURE 5. **The proposed pipeline for 3D object detection and pose estimation.**

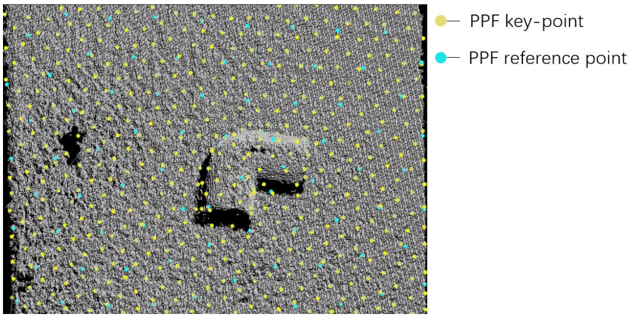

FIGURE 6. **The key point and reference point of scenes, where yellow and blue denote the key point and the reference point denoted under PPF description, respectively.**

camera and box can then have a closer search range than the usual solution, a layer less, a search layer closer from the bottom of box to top area of the target. When there are different type of target, the layer gaps are also different because, under the field of view of the camera, the target of picking is determined by the section above, we know the major workpiece of one layer, the threshold is between the minimum and maximum diameters, and this represents the adaptive threshold.

Because the 3D camera can directly capture depth information and determine the distance between the camera and the actual scene, the adaptive threshold can remove the foreground and background, reduce the point cloud size, and facilitate the speed of extracting features and matching. The three-dimensional model of the object has been used as prior knowledge. For the stacking case which is depicted in Figure 7, suppose we know the 3D camera max height of the field of view, denoted as C, and the height H of the box with the maximum diameter L of the convex enclosure of the recognized target, and the search strategy is followed in the search region of $(C - H) \leq d_1 \leq (C - H + l_1)$ (top-down from the camera view) as shown in Figure 7 and then using the matching algorithm mentioned above to search for the target. In addition, the results corresponding to the number of total layers can be obtained from $n = 1$ and formulated in Equation 3:

$$\begin{aligned} C - H \leq d_1 \leq C - H + l_1, \quad (n = 1) \\ C - H + l_1 \leq d_2 \leq C - H + l_1 + l_2, \quad (n = 2) \\ C - H + l_1 + l_2 \leq d_3 \leq C - H + l_1 + l_2 + l_3, \quad (n = 3) \\ \vdots \end{aligned} \tag{3}$$

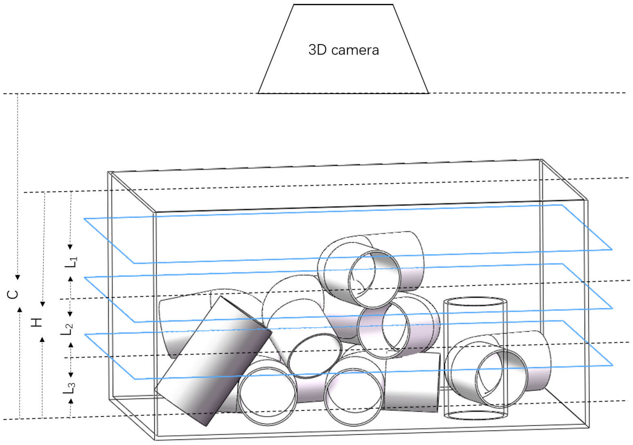

FIGURE 7. **Mathematical model of picking strategy.**

Taking a special situation as an example, if all layer maximum diameter values L of the convex enclosure are the same, i.e., $l_1 = l_2 = l_3 \ldots$, then the above search range can be described as $C - H + (n - 1) * L \leq d \leq C - H + n * L$. The pseudocode statement is shown in Algorithm 1, in which the input includes the distance between box and camera, which is denoted as C, the box height H, maximum workpiece height L_{max}, and predefined minimum similarity score threshold S_{min}. The output is the pose of target P_{final} and the similarity score S_{final}.

As the code shows, we consider the case in which the condition $S > S_{min}$ is satisfied at $n = n_0$. Next, step $n = n_0 - 1$, pose P, score S and the search range ρ will be updated according to the algorithm. If not satisfied, the search range will be updated using the predefined threshold Δ. The side view of our bin-picking multilayer algorithm result is shown in Figure 8. The green coloration and coordinates represent the recognized target using the multilayer adaptive threshold picking strategy described above. The piped *corner-joint* has been masked with green on every recognized object. As shown in Figure 9, the first row is the real scene of the 3D camera, where the target recognized by our method is indicated by a green mask for convenience. If the target has been recognized, the green mask with ICP pose refinement will be used, and the others will retain the original color, i.e., green coloration indicates the recognized objects, whereas the gray objects are not recognized. The second row is the processed point cloud in our experiment, from left to right figures are the targets which were highlighted using our

Algorithm 1: Multilayer Adaptive Threshold Picking Strategy

```
Input:
$H$: Height of the box;
$L$: Height of the workpiece;
$C$: Distance between the box and the camera;
$\rho$: The threshold of each layer;
$\Delta$: A constant value.
$P$: Temporary pose value.
$S$: Temporary pose score.
$S_{min}$: Predefined minimum similarity score threshold.
Result: Target pose $P_{final}$, the similarity score $S_{final}$
Total number of layers: $n = ceil(H/L_{max})$
Minimum threshold: $\rho_{min} = C - H + (n - 1) * L$
Maximum threshold: $\rho_{max} = C - H + n * L$
for $n >= 1$ do
    Pose estimation in the range of $\rho_{min}$ and $\rho_{max}$
    $P \leftarrow$ Updating the resulting pose
    $S \leftarrow$ Updating the resulting similarity score
    if $S > S_{min}$ then
        $n = n - 1$
        $P_{final} \leftarrow P$
        $S_{final} \leftarrow S$
        $\rho_{min} = C - H + (n - 1) * L$
        $\rho_{max} = C - H + n * L$
    else
        $\rho_{min} -= \Delta$
        $\rho_{max} += \Delta$
    end
end
```

method from top to bottom in the box. Here, we use a pass-through and voxel filter to remove the background and noise point cloud.

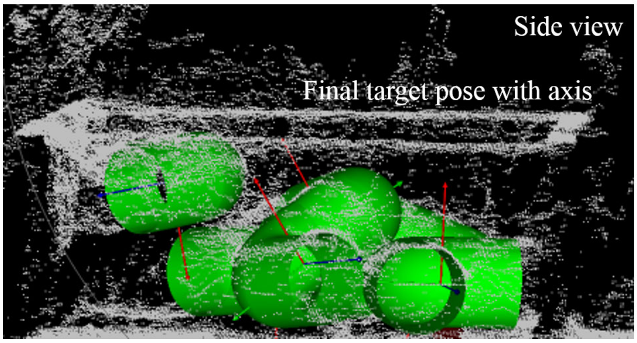

FIGURE 8. A side view image showing the results of the recognized targets in green and the corresponding coordination system using the proposed fast multilayer adaptive threshold.

IV. EXPERIMENTAL VERIFICATION

The setup for the random bin-picking experiment is shown in Figure 10. A laptop with Windows 10 is designed for the system, which has 12GB RAM and an Intel Core i5-4200H CPU (2.8GHz). A 3D industrial depth camera (Ensenso N20-602-16BL) is mounted on top of the bin with appropriate intrinsic and extrinsic parameters calibrated. The robot arm we used for the experiment is the Kinova Jaco^2, which has 6 degrees of freedom, and a payload of 2.6 kg. Combined with an air pump, electromagnetic valves, and relays, we control the relay for the gripper to open and apply suction when gripping the object. In the experiments, 4 different types of objects to be recognized were placed randomly as shown in Figure 11. Here, we recognized 4 different types of objects in total and take the *corner-joint* as an example for further explanation, which is shown in Figure 10. Additionally, supplemental videos demonstrate the whole implementation procedure of the bin-picking, which notice that the design of the multilayer adaptive threshold for picking strategy focuses on increasing the picking speed. However, those situations involving objects under certain poses with insufficient areas exposed for suction, causing picking failures, are outside of the scope of this paper.

As shown in Figure 11, four objects are considered in every scene. The Ground Truth (GT) poses of all objects were manually oriented. Here, we recorded the ground truth of several unstructured objects by manually using a chessboard

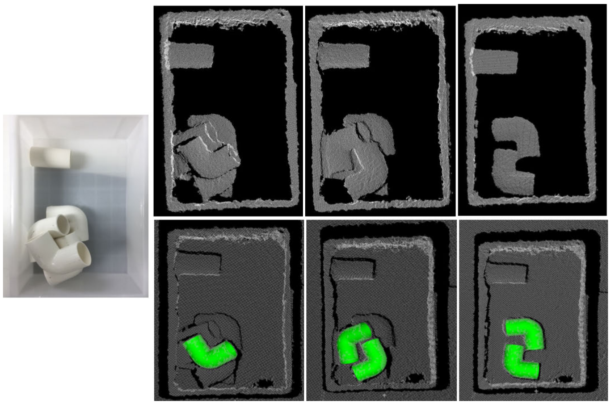

FIGURE 9. Illustrates the results using fast multi-layer adaptive threshold method, where the leftmost image show the real scene, the first row denotes the post-processing point cloud in our experiment as well as the second row represents the recognition targets (highlighted in green for convenience) from left to right and from top layer to bottom.

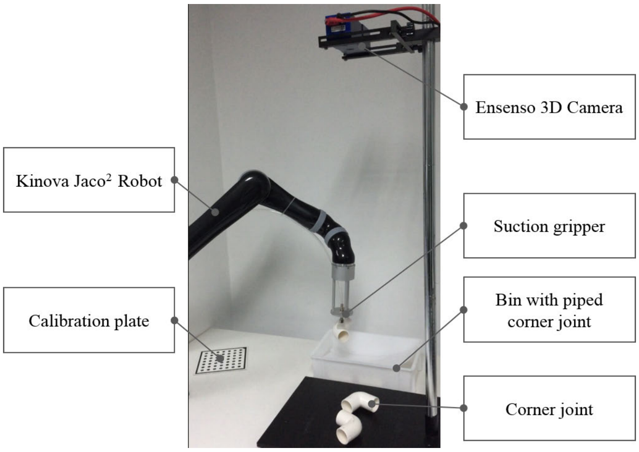

FIGURE 10. Experimental setup for evaluating the proposed bin-picking pipeline using a Kinova robot and 3D depth camera mounted on top of the box.

TABLE 1. Recognition rates of 4 different types of targets with CAD model or real scene models.

Object type	Recognition total times	Failure times	Success rate	Time(unit: s)
Computer-memory-card	30	4	86.67%	0.31
Two-way-joint	32	0	100.00%	0.16
Corner-joint	40	2	95.00%	0.29
Aluminum-motor-support-ring	30	5	83.33%	0.30
Average recognition rates			91.25%	0.265

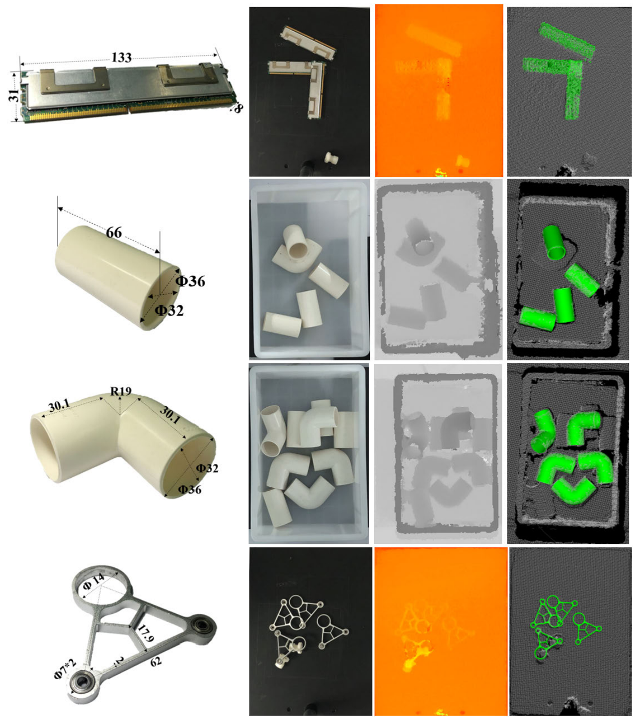

FIGURE 11. Multiple object detection, where the first column is the RGB image (*computer-memory-card, two-way-joint, corner-joint* and *aluminum-motor-support-ring*), the second column is depth image (the first and last depth are highlighted with pseudo color), and the last column is the target with the green pose mask produced using our algorithm.

such that the resolution is on the millimeter scale with respect to the camera frame. We then set the estimated position and orientation as robot input for bin-picking tasks. As depicted in the image, the first column is the RGB image (*computer-memory-card*, *corner-joint*, *two-way-joint* and *aluminum-motor-support-ring* from top to bottom), the second column is the depth image (the first and last depth images are highlighted with pseudo color because they are very thin), and the last column is the target with green pose mask obtained using our algorithm. Once the target is recognized, then a green mask will be attached to the gray point cloud, and the related 6-DoF pose and score are the outputs of our algorithm.

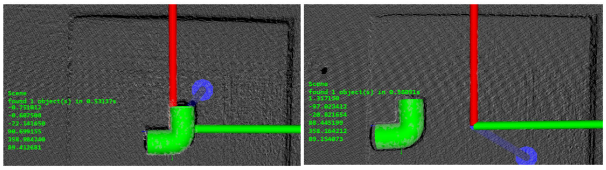

FIGURE 12. Object on origin reference point (left image) and detected result with respect to origin (right image), where the green labels at the bottom of left in the two pictures are the position and orientation of the target.

The resulting four object recognition rates are shown in Table 1, where the *computer-memory-card* is using real scene segmented data as a database model, and *two-way-joint*, *corner-joint* and *aluminum-motor-support-ring* are using CAD-based model data that the first one does not have a bottom point cloud data feature. Hence, the recognition rate of the *computer-memory-card* is slightly lower than the other three. To verify the accuracy of our 6-DoF pose estimation system, the target workpiece selected in the experiment was the *two-way-joint* modeled above; only one single target was placed at one time. In the visible field of view of the camera, the world coordinate system was established with the center of the calibration plate as the reference point; the recognized position and orientation of the workpiece by the camera are transferred to this coordinate system, and the actual length value of the object can be directly measured from the center of the calibration plate, as shown in Figure 12. The following 10 sets of data are randomly placed in 10 different poses, using the proposed methods to acquire the final refined pose.

By conducting approximately 132 picking trials on the four different types of target, *two-way-joint*, *corner-joint* and *aluminum-motor-support-ring* using CAD model data, whereas the *computer-memory-card* uses real scene segmented data, we achieve an average success rate of 91.25% as shown in Table 1. Additionally, we take the *corner-joint* as an example for evaluating the estimation results and its errors of X, Y, Z are shown in Table 2. After calculation, the maximum localization error of the workpiece in the X-axis direction is 1.02 mm, and the root mean square error (RMSE) is 0.79 mm. The maximum localization error of the workpiece in the Y-axis direction is 1.0 mm and the RMSE is 0.85 mm. For Z, the workpiece maximum localization error in that direction is 1.91 mm, and the RMSE is 1.8 mm. Because the 3D depth

TABLE 2. Comparisons of the resulting measurements and ground truth (unit: *mm*).

ID	X				Y				Z			
	RESULT	GT	|ERROR|	RMSE	RESULT	GT	|ERROR|	RMSE	RESULT	GT	|ERROR|	RMSE
1	-0.75	-1.51	0.76	0.79	-0.61	0.22	0.83	0.85	-22.14	-23.81	1.67	1.8
2	1.32	2.13	0.81		-97.02	-97.81	0.79		-20.82	-19.10	1.72	
3	-53.94	-52.92	1.02		-29.99	-29.1	0.89		-21.55	-19.73	1.82	
4	67.17	67.93	0.76		29.02	29.85	0.83		-22.31	-24.22	1.91	
5	-7.77	-8.57	0.80		44.05	43.26	0.79		-21.96	-20.14	1.82	
6	-98.37	-97.84	0.53		-9.91	-9.14	0.77		-21.34	-23.16	1.82	
7	-73.96	-74.6	0.64		-70.32	-71.23	0.91		-20.61	-22.4	1.79	
8	11.66	10.86	0.80		56.70	55.7	1.00		-21.80	-19.95	1.85	
9	78.20	79.02	0.82		5.70	4.82	0.88		-22.43	-20.68	1.75	
10	-71.58	-70.67	0.91		-16.47	-17.28	0.81		-21.08	-22.89	1.81	

camera is sensitive to distance and noise in the Z direction, the error in the Z direction is more substantial than those in the X and Y directions.

V. CONCLUSION AND FUTURE WORK

This paper proposes a CAD-based 6-DoF pose estimation pipeline for robotic random bin-picking tasks using the 3D camera. Less data and time were consumed by using our method for database building and pose estimation. The 3D model database can be generated from either 3D CAD parts or real scene segmented data, which is more convenient in practical production scenarios than the most commonly employed methods. Additionally, as for on-line pose estimation, the improved PPF voting algorithm with multilayer adaptive threshold picking strategy was introduced, and a series of experiments involving practical robotic random bin-picking of multiple object types showed that the proposed system can pick up all randomly posed objects in the bin. A practical experiment in the laboratory achieved an average recognition rate of 91.275 % and a time of 0.265 seconds. For the current system, there is still room to improve the quality of the predicted pose score in clustered scenes, as well as the small size objects, occlusion and better quality of selected suction cup locations on objects. Those will be the topic of our future research.

REFERENCES

[1] G. Lowe, "SIFT-the scale invariant feature transform," *Int. J*, vol. 2, pp. 91–110, Mar. 2004.

[2] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2006, pp. 404–417.

[3] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. ICCV*, Nov. 2011, vol. 11, no. 1, p. 2.

[4] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (FPFH) for 3D registration," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2009, pp. 3212–3217.

[5] R. B. Rusu, "Semantic 3D object maps for everyday manipulation in human living environments," *KI- Künstliche Intell.*, vol. 24, no. 4, pp. 345–348, Nov. 2010.

[6] F. Tombari, S. Salti, and L. Di Stefano, "Unique signatures of histograms for local surface description," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 356–369.

[7] F. Tombari, S. Salti, and L. Di Stefano, "A combined texture-shape descriptor for enhanced 3D feature matching," in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 809–812.

[8] A. Johnson, "Spin-images: A representation for 3-D surface matching," Ph.D. dissertation, Robot. Inst., Carnegie Mellon Univ., Pittsburgh, PA, USA, 1997.

[9] A. Aldoma, M. Vincze, N. Blodow, D. Gossow, S. Gedikli, R. B. Rusu, and G. Bradski, "CAD-model recognition and 6DOF pose estimation using 3D cues," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 585–592.

[10] Z.-C. Marton, D. Pangercic, R. B. Rusu, A. Holzbach, and M. Beetz, "Hierarchical object geometric categorization and appearance classification for mobile manipulation," in *Proc. 10th IEEE-RAS Int. Conf. Hum. Robots*, Dec. 2010, pp. 365–370.

[11] Z.-C. Marton, D. Pangercic, N. Blodow, and M. Beetz, "Combined 2D–3D categorization and classification for multimodal perception systems," *Int. J. Robot. Res.*, vol. 30, no. 11, pp. 1378–1402, Sep. 2011.

[12] P. J. Besl and N. D. McKay, "Method for registration of 3-D shapes," *Proc. SPIE*, vol. 1611, Apr. 1992, pp. 586–606.

[13] S. Rusinkiewicz and M. Levoy, "Efficient variants of the ICP algorithm," in *Proc. 3rd Int. Conf. 3-D Digit. Imag. Model.*, vol. 1, 2001, pp. 145–152.

[14] A. Segal, D. Haehnel, and S. Thrun, "Generalized-ICP," in *Robotics: Science and Systems*, vol. 2, no. 4. Seattle, WA, USA: RSS, 2009, p. 435.

[15] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce, "3D object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints," *Int. J. Comput. Vis.*, vol. 66, no. 3, pp. 231–259, Mar. 2006.

[16] M. Ulrich, C. Wiedemann, and C. Steger, "CAD-based recognition of 3D objects in monocular images," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2009, pp. 1191–1198.

[17] C. Choi and H. I. Christensen, "Real-time 3D model-based tracking using edge and keypoint features for robotic manipulation," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2010, pp. 4048–4055.

[18] A. Collet, M. Martinez, and S. S. Srinivasa, "The MOPED framework: Object recognition and pose estimation for manipulation," *Int. J. Robot. Res.*, vol. 30, no. 10, pp. 1284–1306, Sep. 2011.

[19] S. Salti, A. Petrelli, F. Tombari, and L. Di Stefano, "On the affinity between 3D detectors and descriptors," in *Proc. 2nd Int. Conf. 3D Imag., Modeling, Process., Vis. Transmiss.*, Oct. 2012, pp. 424–431.

[20] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, J. Wan, and N. M. Kwok, "A comprehensive performance evaluation of 3D local feature descriptors," *Int. J. Comput. Vis.*, vol. 116, no. 1, pp. 66–89, Jan. 2016.

[21] K.-T. Song, C.-H. Wu, and S.-Y. Jiang, "CAD-based pose estimation design for random bin picking using a RGB-D camera," *J. Intell. Robotic Syst.*, vol. 87, nos. 3–4, pp. 455–470, Sep. 2017.

[22] S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit, "Multimodal templates for real-time detection of textureless objects in heavily cluttered scenes," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 858–865.

[23] B. Drost, M. Ulrich, N. Navab, and S. Ilic, "Model globally, match locally: Efficient and robust 3D object recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 998–1005.

[24] T. Hodan, F. Michel, E. Brachmann, W. Kehl, A. GlentBuch, D. Kraft, B. Drost, J. Vidal, S. Ihrke, and X. Zabulis, "BOP: Benchmark for 6D object pose estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 19–34.

[25] S. Hinterstoisser, V. Lepetit, N. Rajkumar, and K. Konolige, "Going further with point pair features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 834–848.

[26] D. Liu, S. Arai, J. Miao, J. Kinugawa, Z. Wang, and K. Kosuge, "Point pair feature-based pose estimation with multiple edge appearance models (PPF-MEAM) for robotic bin picking," *Sensors*, vol. 18, no. 8, p. 2719, 2018.

[27] D. Liu, S. Arai, Z. Feng, J. Miao, Y. Xu, J. Kinugawa, and K. Kosuge, "2D object localization based point pair feature for pose estimation," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Dec. 2018, pp. 1119–1124.

[28] J. Vidal, C.-Y. Lin, and R. Marti, "6D pose estimation using an improved method based on point pair features," in *Proc. 4th Int. Conf. Control, Autom. Robot. (ICCAR)*, Apr. 2018, pp. 405–409.

[29] J. Vidal, C.-Y. Lin, X. Lladó, and R. Martí, "A method for 6D pose estimation of free-form rigid objects using point pair features on range data," *Sensors*, vol. 18, no. 8, p. 2678, 2018.

[30] M. Li and K. Hashimoto, "Curve set feature-based robust and fast pose estimation algorithm," *Sensors*, vol. 17, no. 8, p. 1782, 2017.

[31] M. Li and K. Hashimoto, "Accurate object pose estimation using depth only," *Sensors*, vol. 18, no. 4, p. 1045, 2018.

[32] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D ShapeNets: A deep representation for volumetric shapes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1912–1920.

[33] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka, "3D bounding box estimation using deep learning and geometry," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7074–7082.

[34] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 652–660.

[35] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5099–5108.

[36] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes," 2017, *arXiv:1711.00199*. [Online]. Available: http://arxiv.org/abs/1711.00199

[37] M. Rad and V. Lepetit, "BB8: A scalable, accurate, robust to partial occlusion method for predicting the 3D poses of challenging objects without using depth," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3828–3836.

[38] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, "SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1521–1529.

[39] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox, "DeepIM: Deep iterative matching for 6D pose estimation," *Int. J. Comput. Vis.*, vol. 128, no. 3, pp. 657–678, Mar. 2020.

[40] C. Wang, D. Xu, Y. Zhu, R. Martin-Martin, C. Lu, L. Fei-Fei, and S. Savarese, "DenseFusion: 6D object pose estimation by iterative dense fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3343–3352.

[41] M. Fu and W. Zhou, "DeepHMap++: Combined projection grouping and correspondence learning for full DoF pose estimation," *Sensors*, vol. 19, no. 5, p. 1032, 2019.

[42] E. Brachmann, F. Michel, A. Krull, M. Y. Yang, S. Gumhold, and C. Rother, "Uncertainty-driven 6D pose estimation of objects and scenes from a single RGB image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3364–3372.

[43] P. Wohlhart and V. Lepetit, "Learning descriptors for object recognition and 3D pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3109–3118.

[44] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," 2018, *arXiv:1809.10790*. [Online]. Available: http://arxiv.org/abs/1809.10790

[45] 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 *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 5784–5789.

[46] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view RGB-D object dataset," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 1817–1824.

[47] P. Marion, P. R. Florence, L. Manuelli, and R. Tedrake, "Label fusion: A pipeline for generating ground truth labels for real RGBD data of cluttered scenes," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 1–8.

[48] T. Hodan, P. Haluza, S. Obdrzalek, J. Matas, M. Lourakis, and X. Zabulis, "T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 880–888.

[49] C.-H. Wu, S.-Y. Jiang, and K.-T. Song, "CAD-based pose estimation for random bin-picking of multiple objects using a RGB-D camera," in *Proc. 15th Int. Conf. Control, Autom. Syst. (ICCAS)*, Oct. 2015, pp. 1645–1649.

[50] C. Choi, Y. Taguchi, O. Tuzel, M.-Y. Liu, and S. Ramalingam, "Voting-based pose estimation for robotic assembly using a 3D sensor," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 1724–1731.

WU YAN received the M.E. degree in mechanical engineering from the Guangdong University of Technology, in 2016. Since then, he has been a Research Assistant with the Guangdong Institute of Intelligent Manufacturing, Guangzhou, China. His current research interests include 3D perception and robotics.

ZHIHAO XU (Student Member, IEEE) received the B.E. and Ph.D. degrees from the School of Automation, Nanjing University of Science and Technology, Nanjing, China, in 2010 and 2016, respectively. He is currently a Postdoctoral Fellow of the robotics team with the Guangdong Institute of Intelligent Manufacturing, Guangzhou, China. His main research interests include neural networks, force control, and intelligent information processing.

XUEFENG ZHOU received the M.E. and Ph.D. degrees in mechanical engineering from the South China University of Technology, Guangzhou, China, in 2006 and 2011, respectively. He is currently the Director of the robotics team with the Guangdong Institute of Intelligent Manufacturing, Guangzhou. His research interests include neural networks, force control, and intelligent information processing. He received the Best Student Paper Award from the 2010 IEEE International Conference on Robotics and Biomimetics.

QIANXING SU received the B.S. degree in mechanical engineering from Northeastern University at Qinhuangdao, China. Since then, he has been a Research Assistant with the Guangdong Institute of Intelligent Manufacturing, Guangzhou, China. His current research interests include force control and deep reinforcement learning.

SHUAI LI (Member, IEEE) received the B.E. degree in precision mechanical engineering from the Hefei University of Technology, China, in 2005, the M.E. degree in automatic control engineering from the University of Science and Technology of China, China, in 2008, and the Ph.D. degree in electrical and computer engineering from the Stevens Institute of Technology, USA, in 2014. He is currently an Associate Professor (Reader) with the School of Engineering, Swansea University, Swansea, U.K.. His current research interests include dynamic neural networks, robotic networks, and other dynamic problems defined on graphs.

HONGMIN WU received the Ph.D. degree in mechanical engineering from the Guangdong University of Technology, Guangzhou, China, in 2019. He is currently a Postdoctoral Fellow of the robotics team with the Guangdong Institute of Intelligent Manufacturing, Guangzhou. His research interests include in the domain of multimodal time series modeling using Bayesian nonparametric methods, robot introspection, movement representation, and robot learning.