2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Macau, China, November 4-8, 2019

Pose-Aware Placement of Objects with Semantic Labels - Brandname-based Affordance Prediction and Cooperative Dual-Arm Active Manipulation

Yung-Shan Su[1], Shao-Huang Lu[1], Po-Sheng Ser[1], Wei-Ting Hsu[1], Wei-Cheng Lai[1], Biao Xie[2], Hong-Ming Huang[1], Teng-Yok Lee[4], Hung-Wen Chen[5], Lap-Fai Yu[3], Hsueh-Cheng Wang[1,*]

Abstract— **The Amazon Picking Challenge and the Amazon Robotics Challenge have shown significant progress in object picking from a cluttered scene, yet object placement remains challenging. It is useful to have pose-aware placement based on human and machine readable pieces on an object. For example, the *brandname* of an object placed on a shelf should be facing the human customers. The robotic vision challenges in the object placement task: a) the semantics and geometry of the object to be placed need to be analysed jointly; b) and the occlusions among objects in a cluttered scene could make it hard for proper understanding and manipulation. To overcome these challenges, we develop a pose-aware placement approach by spotting the semantic labels (e.g., brandnames) of objects in a cluttered tote and then carrying out a sequence of actions to place the objects on a shelf or on a conveyor with desired poses. Our major contributions include 1) providing an open benchmark dataset of objects and brandnames with multi-view segmentation for training and evaluations; 2) carrying out comprehensive evaluations for our brandname-based fully convolutional network (FCN) that can predict the affordance and grasp to achieve pose-aware placement, whose success rates decrease along with clutters; 3) showing that active manipulation with two cooperative manipulators and grippers can effectively handle the occlusion of brandnames. We analyzed the success rates and discussed the failure cases to provide insights for future applications. All data and benchmarks are available at https://text-pick-n-place.github.io/TextPNP/**

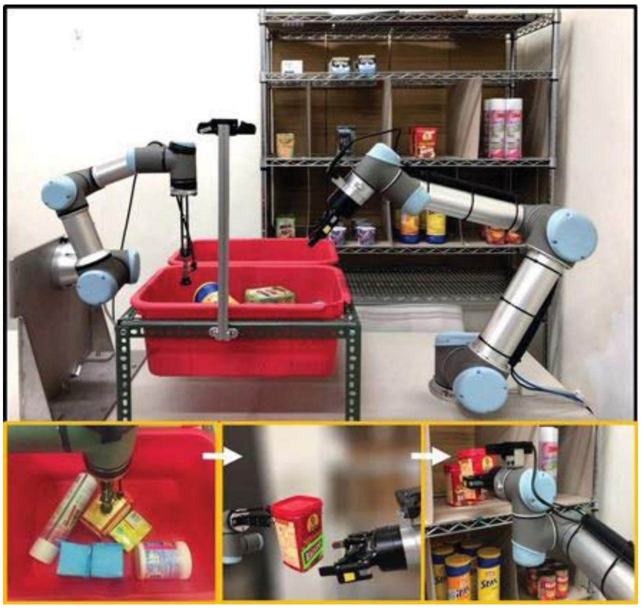


Fig. 1: We use the product brandname, one of the “semantic labels,” for pose-aware placement of objects. We perform active manipulation using two cooperative robotic arms to handle object occlusions (e.g., brandname facing downward). Bottom left to right: A vacuum gripper picks a target object using object-level or brandname-level affordance prediction, the brandname is then used to predict grasp, and finally a two-finger gripper places the target object on the shelf.

I. INTRODUCTION

Robotic pick-and-place systems are in domain in many areas such as servicing, warehouse and grocery store antomation, etc. Recently an increasing number of robots are being used in the factory assembly production lines or warehouses to reduce the human labor involved. The world-class competitions of Amazon Picking challenge and Amazon Robotics Challenge (APC/ARC) 2015-2017 further brought together various teams to develop picking systems for known and unknown objects from the shelves and totes. The common workflow for the picking tasks, especially the solutions for APC/ARC, involve the localization of individual objects via pixel-wise semantic segmentation (e.g., Fully Convolution Networks [1]) or bounding-box-based object detection (e.g., Faster RCNN [2], SSD [3], YOLO [4], etc), the estimation of object poses (e.g., geometric model fitting methods such as iterative closest points [5]), the selection of objects to pick (e.g., estimating the probability of picking success [6]), and grasping the selected objects. Predicting the grasping locations using learning approaches have also been extensively studied [7], [8], [9], [10].

Although picking and grasping predictions have been dealt with in previous studies, the *placement* of objects is still not much addressed in previous work, especially when geometry is not the sole consideration for placing an object. Such scenarios include product stocking, inventory taking, and checkout. When human customers shop in a store or a supermarket, they intuitively pay attention to brandname printed on the product; therefore, the products have to be

The authors are with the [1]Department of Electrical and Computer Engineering, National Chiao Tung University, Taiwan. *Corresponding author email: hchengwang@g2.nctu.edu.tw

[2]Department of Computer Science, University of Massachusetts at Boston, USA

[3]Department of Computer Science, George Mason University, USA

[4]Mitsubishi Electric Research Laboratories, Cambridge, MA, USA

[5]Delta Research Center, Taiwan

978-1-7281-4004-9/19/$31.00 ©2019 IEEE

placed on the shelves with the brandname facing outward. As a result, the estimates of the semantics and the geometry of the object for pose-aware placement are important to create practical values.

We refer to a *semantic label* as a piece of surface on an object that provides not only geometry, but also additional information to facilitate manipulations, such as the references to fulfill a task-relevant object pose, better object duplicates handling, or inferring how to perform proper sequence of actions. There may be more than one semantic label on an object, such as one of the six faces of a cubic object, a soda can lid, and so on. This study focuses on placement task based on the brandname semantic labels.

In this paper, we present an end-to-end pose-aware placement approach that uses the brandname as the reference for object pose and the affordance of the grippers. Our contributions are as follows:

a) Brandname-based Affordance Prediction: Brandname is one of the semantic labels on an object, and it exists on almost every commercial product. Brandname is printed on flat surfaces of box containers or the curved surfaces of cylinders, and it is bounded by a rectangle box. Using this information, we can use the visibility of brandname to directly predict the object affordance (i.e., the probability of picking success) for a vacuum gripper or the grasp for two-finger gripper.

b) Active Manipulation with Actions of Two Cooperative Arms and Grippers: Besides *passively* using the result of object detection and pose estimation, we present schemes to *actively* manipulate the captured images to maximize the visibility of brandnames and the individual objects. If the brandname is invisible, single arm will not be benefited by assembling both two-finger gripper and vacuum gripper since flip or re-grasp may be needed. However, with the setup of dual-arm, the vacuum gripper can lift the product and then the two-finger gripper can observe it. In this case, it can ensure that the camera could perceive the brandname properly. Consequently, there is a great deal of difference between the efficiencies of the two schemes. Our proposed approach can not only change the camera viewpoint to see the brandname with least occlusion, but it can also cooperate with multiple robotic arms and grippers to manipulate the object and achieve the desired placement.

c) Benchmark Dataset With Semantic Labels: Despite the recent progress of object segmentation, training a deep convolutional neural network usually requires a huge dataset of labeled training data. Although there have been attempts to handle the constraints of novel objects and limited data, the progress is still restricted. In our dataset, we select 20 products; our selection is based on the following criteria: 1) the brandname region is at least 1.5 centimeters high, 2) the brandname was on a single line, and 3) the brandname instance occurred only once on the object. To train the vision algorithms for semantic labels, we construct a dataset that included over 8,000 manually-labeled images with brandnames. The dataset includes the training data of real and virtual environments, and a physical benchmark test set for carrying out placement tasks. The datasets are made publicly available [11].

The remainder of the paper is organized as follows. Section II describes the related work on recent advances of affordance prediction as well as active vision and manipulation. We describe the proposed cooperative dual-arm system in Section III, and we show how the baseline and active manipulation are performed in Section V. Section VI describes the "Brandname Benchmark Dataset" including training data from real and virtual environments and a test set with clutter for evaluations. Section VII provides extensive experiments for the proposed methods on the datasets. Finally, we discuss our future work in Section VIII.

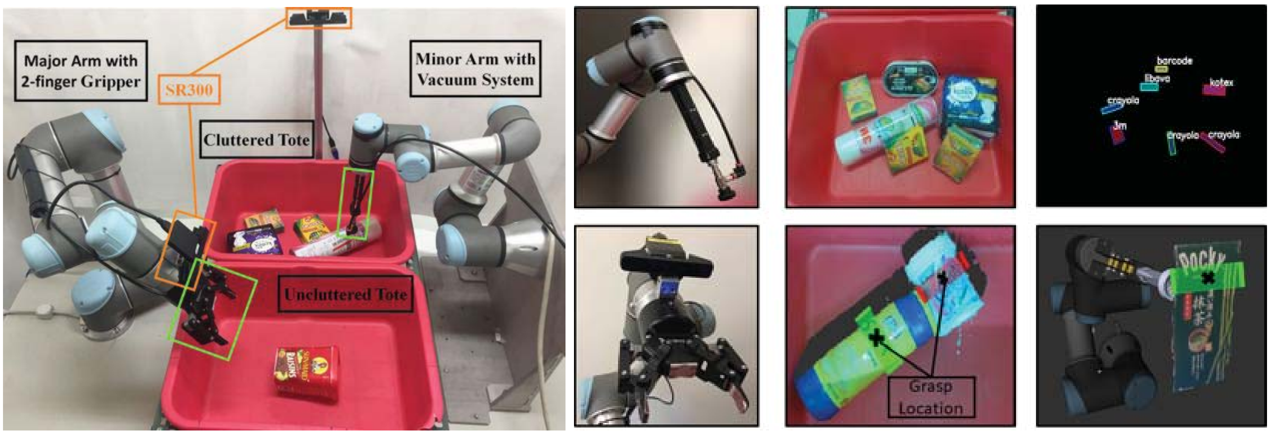
II. RELATED WORK

A. Active Vision and Manipulation

Robotic vision/manipulation, different from the works in computer vision community, has the potentials to control cameras or even manipulate with the scene to improve perception [12]. The topics adopted in [13] [14] [15] used the next-best viewpoint to improve the perception confidence for object detection. [6] follows grasp-first-then-recognize strategy to improve perception in cluttered environments, and it further re-orders objects to enable the objects to be easily grasped by a two-finger gripper. Such advantages help overcome the challenge of pick and place in an occluded and cluttered environment, and they change the scene into a simpler (uncluttered) environment to obtain higher perception.

B. Affordance and Grasp Predictions

Object affordance is an important topic for pick-and-place systems, and the algorithms tend to be highly related with end-effector co-designs. To handle clutter, occlusion conditions, and different object geometries, many recent studies have adopted different affordance predictions together with a customized end-effector: [16] relies on the classic model-based pose-estimation with object registration and the corresponding affordance modes. In [6], the authors have defined four primitives for grasping and suction, and they trained two fully convolutional network (FCN) models to predict the dense pixel-wise affordance probability. Moreover, the affordance, or more precisely, the grasp prediction helps two-finger gripper to execute picking tasks in cluttered environments. In [7], the authors encoded a raw RGB-D image input into several grid cells, and they predicted direct regression to grasp coordinates under the assumption that there was only a single correct grasp per image. Their revised model further predicted multiple grasps per object in real time by using locally constrained predictions. On the other hand, relying on the geometry of object without color information, [9] takes grasp candidates which are aligned to the depth image as inputs and predict the probability of grasp success. In [17], they adopted a multi-stage learning approach that combined convolutional neural network and reinforcement learning to learn the grasping pose. Subsequently, used two FCNs that mapped camera inputs to actions: one FCN pushed with an end-effector orientation



(a) Collaborative robotic arms for pose-aware placing.

(b) Brandname-based affordance and grasp predictions.

Fig. 2: Left: we propose a dual-arm cooperative approach to *actively* manipulate the scene to improve the perception for the placement of objects. The vacuum gripper moves an occluding object to reveal information (the invisible brandname) hidden underneath, and the two-finger gripper further predicts the grasp using the brandname and completes the pose-aware placement. Right: brandname can be used to predict both the affordance for vacuum gripper and the grasp for two-finger gripper.

and location, and the other was used for grasping, and the work in [18] used reinforcement learning to decide whether to push or separate adjacent objects or to pick up objects. All these showed significant progress in solving the picking problem in cluttered environments, but they did not consider placements using desired poses.

C. Dual-Arm Manipulation

Although manipulation problems have been widely studied with single arm, very few studies have investigated dual-arm settings. As shown in the survey by Smith *et al.* [19], different approaches have been introduced in order to distinguish among no coordinations, goal-coordinated approaches (which solves the same task without physical interaction), and bimanual manipulations (which physically interacts with the same object). In [20], the researchers demonstrated how to execute cluttered picking tasks by using dual arm in the goal-coordinated approach; their system showed only how to coordinate in the same working space without any interactions. In [21], the authors have dealt with pose-aware placement problems by considering the geometry of both the limited placement space and the target objects. It searches multiple placement postures and carries out re-grasping object with dual arm, for the purpose of flipping objects and making objects easy to pick in a specific pose. In [22], the authors designed a dual-arm system with two functions; one arm swept objects using a plate, and another lifted the objects using a suction cup in a cluttered container. Our approach can be considered as bimanual, given that the two arms interact while handling a product placed with the brandname facing downward. One vacuum gripper lifts the object so that the other two-finger gripper can pick up and place the object.

III. Approach

To achieve pose-aware placement in a cluttered environment, we developed a dual-arm cooperative approach guided by brandname-based pose-aware affordance prediction, see Fig. 2.

A. System Overview

1) Two Cooperative Arms and Workspace Settings: We fabricate a pose-aware placement system built using two cooperative manipulators (Fig. 2a). One manipulator is a Universal Robotics UR5 equipped with a Robotiq two-finger end effector along with an Intel RealSense SR300 RGB-D camera; the other is a Universal Robotics UR3 with a vacuum gripper, that is used for active manipulation (see Section V) and handling occlusions in clutter. We mount the two manipulators on the same desktop 100 cm from each other. Two totes are placed between the two manipulators: one tote is cluttered and another is not. The workflow starts with the clutter tote. The manipulators may place objects in the uncluttered tote as an intermediate region, or they may place objects directly on a shelf with 6 bins as the final placement positions.

2) Multi-View Active Vision: Our vision system consists of two SR300 RGB-D cameras: one is integrated on the UR5 manipulator, making it possible to deal with the occlusions among objects by controlling the arm and changing the viewpoints. Another camera is mounted on top of the desk facing the cluttered tote. The depth ranges of both RGB-D cameras are from 0.2 m to 1.5 m.

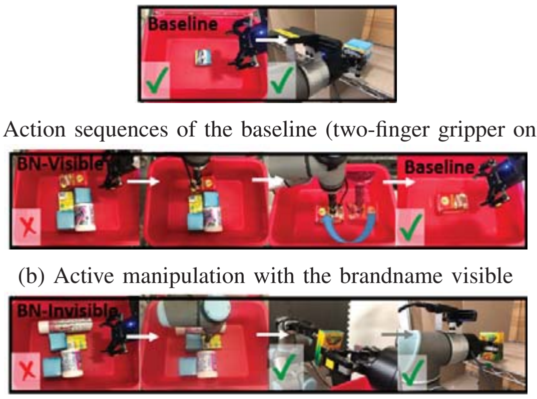
3) Vacuum Gripper for Picking and Two-Finger Gripper for Placing: To place objects in the bin with brandname facing outside, we apply pose-aware affordance to pick objects using the two-finger gripper, see Fig. 2b. However, it's nearly impossible to do this in cluttered environment because of occlusion. Thus, we adopt a two-stage picking and placing with two grippers having different functions. The vacuum gripper first pick objects from the clutter to uncluttered environment, and two-finger gripper pick objects again and pose-aware placement.

IV. Brandname-based Affordance Prediction

Our dual arm system is guided by the affordance map which helps determine how to pick objects from an uncluttered environment using vacuum gripper, and how to pick-and-place objects in the designated bin. The affordance map is based on the rotation-variant brandname segmentation.

1) Brandname Segmentation: Many established object-based detectors generate bounding boxes around targets in a rotation-invariant fashion, but such results are not sufficient to complete the pose-aware placement. We use brandname as the reference for defining object pose, i.e., the surface normal of the brandname region. We want to detect a brandname only when its angle between horizontal line ranges from -45° to 45°, for example, we do not want the brandname is upside down. We train an FCN model using the training set (described in Section VI) with the model parameters that are initialized using the VGG-DICTNET [23]. VGG-DICTNET is a convolutional neural network trained from 8 million computer-graphics rendered training data, and it is able to recognize 88,172 dictionary words used for text classification. We modified the fully connected layers into fully convolutional layers by adding the convolution and deconvolution layers. Finally, the layers turn into the schema of the FCN [1], whick results in VGG-DICTNET-FCN. The VGG-DICTNET-FCN model downsamples the input image by a factor of 32 and then upsamples the image to the original size. The VGG-DICTNET-FCN model takes a grayscale image as the input and returns a set of 21 densely labeled pixel probability maps, which includes 20 maps for the brandnames and one for the background. Then, we run this prediction model four times, by rotating at 0°, 90°, 180°, and 270°, on a single camera view, and we expect to see the brandname segmentation only in one out of four predictions. Then, we combined these four predictions into a mask called affordance map by rotating each prediction to its original direction respectively. The connected pixels of the segmentations are then clustered as an image mask that covered entire brandname. Sometimes, the brandname pixels occur in two of the predictions when their rotation is close to -45° or 45°, and we will choose the result with a larger mask. Finally, the masks from the prediction results are used for affordance prediction.

2) Affordance Prediction: By assuming that the brandname can be fitted into a rectangle, we calculate an affordance map based on the mask of the segmentation. Further, we estimate the bounding rectangles to determine the brandname pose based on the aspect ratio, in which we assume that the width of the brand-name polygon is more than its height. The first row of Fig. 2b demonstrates an affordance map predicted from a cluttered tote having six objects. For picking with two-finger gripper, we further estimate the grasp locations of the two-finger gripper based on the predicted brandname bounding rectangles in the affordance map. For picking with the vacuum gripper, we search and filter the area in the affordance map where the surface normal is vertical to the tote and the curvature is lower than the threshold as the final picking point.



(a) Action sequences of the baseline (two-finger gripper only)

(b) Active manipulation with the brandname visible

(c) Active manipulation when the brandname is not visible

Fig. 3: The action sequences are determined by brandname affordance and grasp prediction that trigger the perception-driven finite state machine.

V. Active Manipulation

In a typical pick-and-place case, the occlusions and clutters are the challenges that cause failure. The proposed brandname-based methods also suffer from self-occlusion or occlusions with the other objects. We introduce our baseline, that is, the brandname-based pose-aware placement with two-finger configuration to show that our method could achieve a high success rate for pick-and-place in an uncluttered scene, but would fail if the other objects are in the neighborhood of the target object in clutters. We further propose a two-stage *active* manipulation system to overcome the brandname occlusion even for cases when the brandname is hidden underneath.

A. Baseline: Two-Finger Gripper Pick and Place

The baseline solution is capable of handling the cases when the brandname on the object is visible in an uncluttered scene: the two-finger gripper is a good way of executing pose-aware placement tasks because of the stability. The grasp is estimated using the brandname segmentation pipeline; we assume that the transforms from the brandname poses to grasp poses are known for each object and its brandname (Fig. 3a).

B. Active Manipulation

1) BN Visible: Vacuum Pick-n-place, Two-Finger Pick-n-Place: There are still challenges to successfully pick up an object from a cluttered tote even when the brandname is visible because there could be potential collisions with other objects. Previous works in APC/ARC have shown that the vacuum gripper outperforms the two-finger gripper in the picking stage. Thus, we adopt the strategy of using a vacuum gripper to pick an object and place it in the center of the other uncluttered tote. Then, we conduct another pick-and-place with two-finger gripper (Fig. 3b). The first pick-and-place also includes rotating the object so that brandname can be easily predicted at the desired degree.

2) BN Invisible: Vacuum Pick, Two-Finger Pick-n-Place: If the brandname is invisible, the objects need to be lifted, in order to obtain the brandname underneath. Then, we perform the second pick-and-place by using a two-finger gripper to reach the designated shelf (Fig. 3c). However, during the first stage we can only rely on the object-level FCN. Affordance prediction is based on object segmentations and the surface normal of the point cloud to determine a picking point where there should not be object edges or boundaries.

VI. The Brandname Benchmark Dataset

To the best of our knowledge, our Brandname Benchmark Dataset is the first dataset that targets pose-aware placement using semantic labels. There are some relevant works such as the scene-level Grocery Dataset [24] that collects 25 classes of objects, and targets classification problem instead of segmentation, or object-level "Shelf & Tote" Benchmark Dataset [25]. Our datasets include object-level and brandname-level annotations for the following three collections: 1) A training set from real environment with image variances of the illumination changes, object reflectiveness, and different tint. 2) A virtual training set collected within a simulation environment that follows the real environment settings, and the ground truth of objects and brandnames (or other semantic labels such as barcodes) are automatically annotated, see Fig. 4 and Table I. Both 1) and 2) contain only one object in a scene based on the findings of the batch-training (i.e., training on object A only or object B only enables successful prediction on the images with both objects A and B) [26], [25]. 3) A test set containing scenes in which there are multiple objects in clutter, either duplicated or multiple objects, and a certain level of occlusions. The test set will be used to carry out the baseline and active manipulations in real environments, as shown in Fig. 5.

A. Training Sets in Real Environments

The data collection follows the method given in [25] which appeared in the Amazon Picking Challenge 2016. In each scene, a single object is placed in the tote with an arbitrary initial pose. An RGB-D camera mounted on the UR5 is used to systematically capture images automatically in multiple *views*. There are 920 scenes × 31 views. There are 22 missing images from the total of 28,520 images because of hardware problems, which resulted in 28,498 images.

1) Objects: For the pixel-wise object label generation, we use FCN and the image processing algorithm to construct a semi-automatic data-labeling system. In our object segmentation scenario, the image from the camera would contain only a single object and tote in one scene; therefore, only one tote would be labeled as background. Because of the monotone of the background, a two-class output classifier model (one class is the object and the other class is background) that is trained on a small dataset has the ability to distinguish the background and the objects. For building our semi-automatic data-labeling system, we starts by manually labeling a small part of our original 500 images using the annotation tool LabelMe [27]. Then, we fine tune the VGG16s-FCN network to the two-class output classifier model using the hand-labeled images as training dataset to predict the pixel-wise object segmentation of the remaining pictures. The prediction result might have scatter noises; therefore, we filter the largest area of each object class as the final label. Although the prediction results might not be identical to the manual labels, the approximations with a certain level of high intersection over union (IOU) are able to allow the robotic arms to complete the tasks.

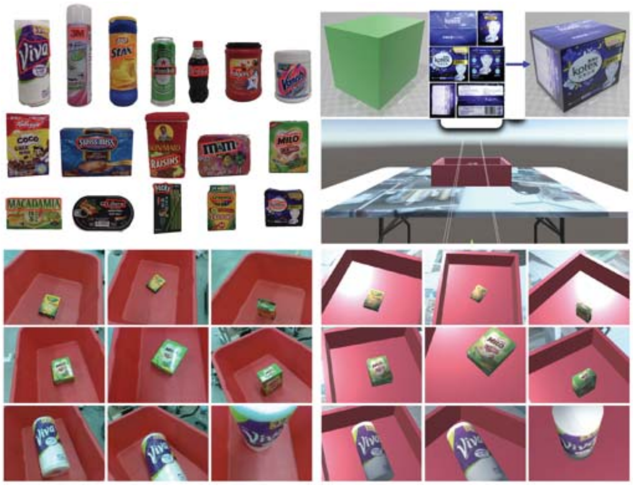
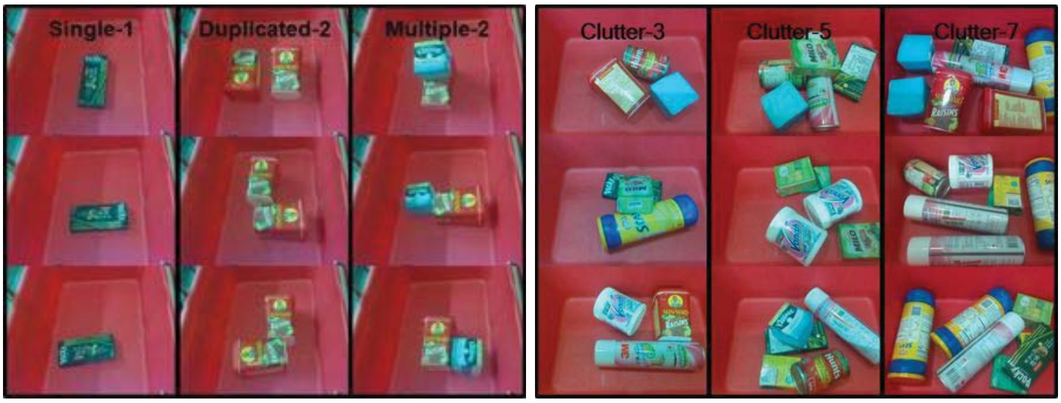


Fig. 4: Data collected for the training data in real (left) and virtual (right) environments. Upper-left: Samples of objects and their brandnames (red polygons). Top-right: Creating an object model via the 3D builder in Unity. Each scene contains only one object, because batch-training using a single object could yield deep models that give good inference results for scenes with multiple objects [26], [25].

2) Brandnames: To train a rotation-variant brandname detector, the predicted brandname is expected to be above a certain IoU. Thus, we design a rotation-variant label criteria for the brandname; the brandname is labeled only if its angle with horizontal line is from -45° to 45° and at least approximately 50% or more of the brandname is visible. Using these criteria, we obtained approximately 30% of the 28,498 images. The brandname areas are labeled as polygons.

B. Training Sets in Virtual Environments

1) Building Virtual Environment: To create virtual environment as similar as possible to real world, we build a tote with similar configuration in Unity and randomly adjust the hue of the light during data collection, see Fig. 4. For the object model, to avoid distortions and have a high resolution, it is manually created in CAD software 3D Builder [28]



(a) Scenes are designed with single, 2 duplicated, or 2 different objects, adjacent or occluded with each other. All brandnames are facing up.

(b) Scenes are arranged with 3, 5, or 7 objects in clutter. The brandnames are either facing upward or downward, and may be occluded.

Fig. 5: Physical benchmark test set is designed with a certain scenes ranging from 1 to 7 objects.

TABLE I: Numbers of scene, view, and data augmentations carried out in the proposed training sets from real and virtual environments. OBJ (object), BN (brandname), and BAR (barcode) are manually annotated in real set and automatically generated in the virtual set.

	Scene	View	Aug.	OBJ	BN	BAR
Real Env.	920	31	-	28,498	8,576	5,686
Virtual Env.	200	54	4	43,200	24,624	14,508

and its texture is imported by six faces of high-resolution images from real object. Those CAD models are then labeled with brandname and barcode for each object. Given these settings, we can efficiently and automatically collect RGB image, object, brandname, and barcode labels.

2) Data Augmentation: Images captured in real environments are often degraded by motion blur and out-of-focus Gaussian noises. Therefore, we add these noises in the collected virtual data to improve the varieties of our virtual dataset. We include two levels of both motion blur and out-of-focus, which resulted in four times more data.

C. Physical Benchmark Test Sets

Unlike the training sets, the test set is designed with 1 to 7 objects in a scene, and some objects may be adjacent or occluded to others. There are six subsets of the physical benchmark test set, see Fig. 5. All brandnames faced upward in the Single-1, Duplicate-2, and Multiple-2 scenarios, but some are invisible because of occlusions. There are 20 scenes with various object placements and occlusions among multiple objects in Clutter-3, Clutter-5, and Clutter-7 subsets. The brandnames are either facing upward or downward, and may be occluded. There are 290 scenes, 710 objects, and 476 visible brandnames manually annotated for evaluations.

VII. Experiments

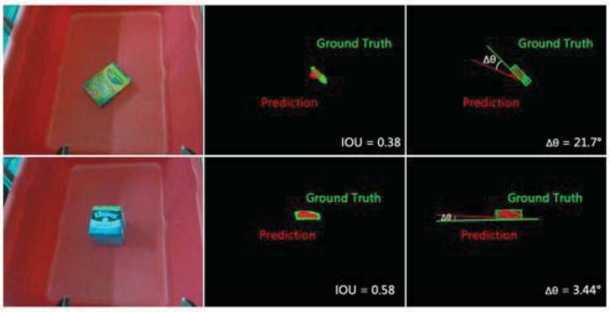


Fig. 6: Samples of IoU and $\Delta\theta$ of brandname segmentations. A High perceptage of low IoU cases and more $\Delta\theta$ may cause more incorrect affordance grasp predictions to successfully compete pose-aware placing.

The physical benchmark dataset is then used for two evaluations: 1) To determine how well the deep models of the brandname segmentation trained from batch-training (i.e., only one object in each training sample) predict the scenes containing multiple objects in clutters. 2) to determine how the predicted affordance and grasp work for end-to-end pose-aware placing for the proposed dual-arm manipulators as compared with the baseline method.

A. Brandname Segmentations

We first evaluate the rotation-variant predictions of brandname segmentations in the image-level and the brandname-level, as shown in Table II. Our metrics include image-level average F-scores $(2 \cdot \frac{precision \cdot recall}{precision + recall})$, brandname-level IoU calculated in pixels, and $\Delta\theta$ (degree) between the predicted and ground-truth rotated rectangles, as shown in Fig. 6. Although the findings in [26] suggest that batch-training with single object generally works well for multi-object scene prediction, we found that the increasing clutters (i.e., number of objects in tote) result in lowering the image-level F-score. Clutters also increase the percentage of low IoU cases and $\Delta\theta$, which indicates low confidences of subsequent

TABLE II: Evaluation of brandname segmentation. Scene: Number of scenes, Vis. BN: Number of visible brandnames, Num.: Number and percentage of brandnames that have IoU < 0.5.

	Image-level		Brandname-level			
Benchmark	Scene	F-score	Vis. BN	IoU < 0.5	IoU ≥ 0.5	
				Num. (%)	Ave. IoU	$\Delta\theta$
Single-1	50	0.70	50	7 (14%)	0.72	5.45
Duplicated-2	90	0.66	145	32 (22%)	0.71	5.91
Multiple-2	90	0.66	159	36 (23%)	0.70	5.64
Clutter-3	20	0.62	31	7 (23%)	0.73	7.14
Clutter-5	20	0.60	32	11 (34%)	0.66	7.77
Clutter-7	20	0.53	59	17 (29%)	0.70	7.90

affordance and grasp predictions.

B. End-to-End Pose-Aware Placement

We first evaluate the baseline solution for brandname visible objects in the subsets of the Physical Benchmark Test Sets. The metrics for the performance of the baseline method include the following:

- Pick Succ.: The picking stage is success if the robot can grasp the object without dropping it before placing it.
- Place Succ.: The placement stage is success if the object is put in the designated bin with brandname facing outward.

We found that the grasp predictions of the brandname segmentations could have a high picking success rate of 0.92, and the overall end-to-end placement success rate of 0.88 in an uncluttered scene. These results are comparable with the success rates in literatures, such as [9]. Nevertheless, when the number of objects increase, the success rates drop, especially in the Clutter-3, Clutter-5, and Clutter-7 subsets, where picking with two-finger gripper does not seem feasible. The baseline solution is not able to deal with BN-DOWN cases with a single arm and gripper.

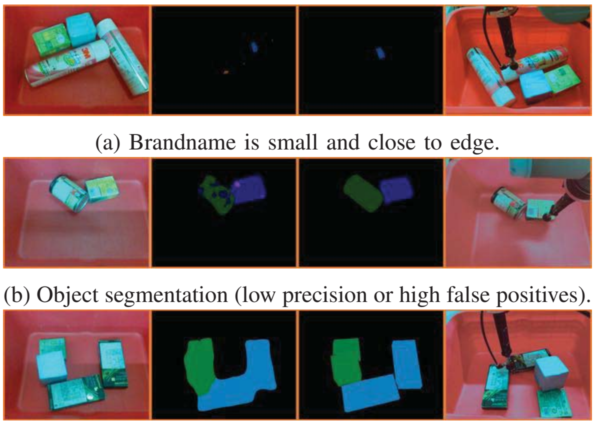
We then evaluate active manipulation, which retrieve objects from the clutters in the first picking; subsequently, we try the second pick and place. Thus we further use the metrics "First Pick Succ.," if the suction cup stably suck the item without dropping it. The "Second Pick Succ." follows the baseline "Pick Succ.". We found that the success rate of the first pick reaches 0.85 and above in the BN-DOWN and BN-UP cases in Clutter-3, and decrease in Clutter-5 and Clutter-7. Active manipulation is able to handle BN-invisible cases that require cooperative dual-arm. Nevertheless, being given a visible brandname does not guarantee that the gripper can grasp the object in cluttered environment because of occlusion. Active manipulation also shows better results than baseline, in which the success rate of placing objects decrease by 34%, 23%, and 24% in Clutter-3, Clutter-5 and Clutter-7 conditions respectively.

Some common failure cases of vacuum gripper are as follows (see Fig. 7):

- The vacuum system tends to fail when the affordance prediction is close to the unflattened surfaces of cylinder (Fig. 7a), and edge of cuboid objects.
- Poor segmentation results affect the affordance prediction, which leads to slight shifts or rotations in the picking affordance. Such cases tend to fail during the placement stage although the picking stage might work.
- If the duplicated objects are adjacent to one another while brandnames are invisible, it would be difficult for the object-level FCN to split the two, see Fig. 7c. The affordance prediction tends to find the adjacent parts of the two items, which makes our vacuum system fail to pick them up.

TABLE III: Evaluations of end-to-end placement. All brandnames of Single-1, Duplicated-2, and Multiple-3 are facing upward (BN-UP). In the other subsets, BN-DOWN represents objects with brandname facing downward. We found that baseline could perform well in an unclutter scene but the performance is severely affected by clutter. Active manipulation shows the capability of handling BN-DOWN cases and retrieving objects from clutters for later placements.

		Baseline		Active		
	Trials	Pick Succ.	Place Succ.	First Pick Succ.	Second Pick Succ.	Place Succ.
Single-1	50	**0.92**	**0.88**	-	-	-
Duplicated-2	180	0.82	0.76	-	-	-
Multiple-2	180	0.81	0.69	-	-	-
Clutter-3	60					
BN-UP	33	0.48	**0.39**	0.88	0.82	**0.73**
BN-DOWN	27	-	-	0.85	0.59	0.59
Clutter-5	100					
BN-UP	35	0.31	**0.26**	0.80	0.57	**0.49**
BN-DOWN	65	-	-	0.75	0.45	0.37
Clutter-7	140					
BN-UP	74	0.23	**0.19**	0.74	0.49	**0.43**
BN-DOWN	66	-	-	0.74	0.38	0.33



(a) Brandname is small and close to edge.

(b) Object segmentation (low precision or high false positives).

(c) Duplicated objects with brandnames facing downward.

Fig. 7: Common failure cases of vacuum gripper. From left to right: Input image, brandname or object segmentations, ground-truth labels, and vacuum gripper actions.

VIII. CONCLUSIONS

This research uses the properties of one of the semantic labels (i.e., brandname) to predict brandname-based affordance and grasp complete pose-aware placement tasks in cluttered environments. We show that dual-arm active manipulation enables robots to retrieve information even from underneath occlusions, regardless of whether the brandname faces upward or downward. Such affordance and grasp predictions are driven by deep models trained in our well-labeled training sets and benchmark test sets. Our comprehensive evaluations suggest future works that might be improved using the proposed virtual datasets. Although the abundant virtual data and their automatically annotated labels create the opportunity to be scalable to a large number of products in real-world store, many automatically generated brandnames may be too small or occluded to be suitable for converging the model training. Finally, from our experiments we found that batch-training may not be ideal to deal with predictions in heavy clutters. It is possible to generate training sets with multiple annotated objects, which may improve affordance and grasp predictions in cluttered environments.

ACKNOWLEDGMENTS

The research was supported by Ministry of Science and Technology, Taiwan (grant 107-2923-E009-004-MY3, 108-2221-E-009-121), and Delta Electronics. We thank Michael S. Solah for his help on our demo video.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[2] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in *Advances in neural information processing systems*, 2015, pp. 91–99.

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “SSD: Single shot multibox detector,” in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[5] F. Pomerleau, F. Colas, R. Siegwart, and S. Magnenat, “Comparing icp variants on real-world data sets,” *Autonomous Robots*, vol. 34, no. 3, pp. 133–148, 2013.

[6] A. Zeng, S. Song, K.-T. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo, N. Fazeli, F. Alet, N. C. Dafle, R. Holladay, I. Morona, P. Q. Nair, D. Green, I. Taylor, W. Liu, T. Funkhouser, and A. Rodriguez, “Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching,” in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2018.

[7] J. Redmon and A. Angelova, “Real-time grasp detection using convolutional neural networks,” in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 1316–1322.

[8] J. Mahler, F. T. Pokorny, B. Hou, M. Roderick, M. Laskey, M. Aubry, K. Kohlhoff, T. Kröger, J. Kuffner, and K. Goldberg, “Dex-net 1.0: A cloud-based network of 3d objects for robust grasp planning using a multi-armed bandit model with correlated rewards,” in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 1957–1964.

[9] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, “Dex-Net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics,” *arXiv preprint arXiv:1703.09312*, 2017.

[10] J. Mahler, M. Matl, X. Liu, A. Li, D. Gealy, and K. Goldberg, “Dex-net 3.0: Computing robust robot vacuum suction grasp targets in point clouds using a new analytic model and deep learning,” *arXiv preprint arXiv:1709.06670*, 2017.

[11] NCTU mobile manipulation 2019. [Online]. Available: https://text-pick-n-place.github.io/TextPNP/

[12] N. Sünderhauf, O. Brock, W. Scheirer, R. Hadsell, D. Fox, J. Leitner, B. Upcroft, P. Abbeel, W. Burgard, M. Milford, *et al.*, “The limits and potentials of deep learning for robotics,” *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 405–420, 2018.

[13] N. Atanasov, B. Sankaran, J. Le Ny, G. J. Pappas, and K. Daniilidis, “Nonmyopic view planning for active object classification and pose estimation,” *IEEE Transactions on Robotics*, vol. 30, no. 5, pp. 1078–1090, 2014.

[14] A. Doumanoglou, R. Kouskouridas, S. Malassiotis, and T.-K. Kim, “Recovering 6d object pose and predicting next-best-view in the crowd,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3583–3592.

[15] M. Malmir, K. Sikka, D. Forster, I. Fasel, J. R. Movellan, and G. W. Cottrell, “Deep active object recognition by joint label and action prediction,” *Computer Vision and Image Understanding*, vol. 156, pp. 128–137, 2017.

[16] C. Hernandez, M. Bharatheesha, W. Ko, H. Gaiser, J. Tan, K. van Deurzen, M. de Vries, B. Van Mil, J. van Egmond, R. Burger, *et al.*, “Team delfts robot winner of the amazon picking challenge 2016,” in *Robot World Cup*. Springer, 2016, pp. 613–624.

[17] L. Pinto and A. Gupta, “Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours,” in *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, 2016, pp. 3406–3413.

[18] A. Zeng, S. Song, S. Welker, J. Lee, A. Rodriguez, and T. Funkhouser, “Learning synergies between pushing and grasping with self-supervised deep reinforcement learning,” *arXiv preprint arXiv:1803.09956*, 2018.

[19] C. Smith, Y. Karayiannidis, L. Nalpantidis, X. Gratal, P. Qi, D. V. Dimarogonas, and D. Kragic, “Dual arm manipulationa survey,” *Robotics and Autonomous systems*, vol. 60, no. 10, pp. 1340–1353, 2012.

[20] M. Schwarz, C. Lenz, G. M. García, S. Koo, A. S. Periyasamy, M. Schreiber, and S. Behnke, “Fast object learning and dual-arm coordination for cluttered stowing, picking, and packing,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 3347–3354.

[21] K. Harada, T. Foissotte, T. Tsuji, K. Nagata, N. Yamanobe, A. Nakamura, and Y. Kawai, “Pick and place planning for dual-arm manipulators,” in *2012 IEEE International Conference on Robotics and Automation*. IEEE, 2012, pp. 2281–2286.

[22] W. Miyazaki and J. Miura, “Object placement estimation with occlusions and planning of robotic handling strategies,” in *2017 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*. IEEE, 2017, pp. 602–607.

[23] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, “Synthetic data and artificial neural networks for natural scene text recognition,” *arXiv preprint arXiv:1406.2227*, 2014.

[24] P. Jund, N. Abdo, A. Eitel, and W. Burgard, “The freiburg groceries dataset,” vol. abs/1611.05799, 2016. [Online]. Available: https://arxiv.org/abs/1611.05799

[25] A. Zeng, K.-T. Yu, S. Song, D. Suo, E. Walker Jr, A. Rodriguez, and J. Xiao, “Multi-view self-supervised deep learning for 6-D pose estimation in the amazon picking challenge,” in *ICRA*, 2017.

[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[27] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, “Labelme: a database and web-based tool for image annotation,” *International journal of computer vision*, vol. 77, no. 1-3, pp. 157–173, 2008.

[28] 3D Builder. [Online]. Available: https://www.microsoft.com/zh-tw/p/3d-builder/9wzdncrfj3t6?activetab=pivot:overviewtab