Reasoning Pose-aware Placing 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/Robotics Challenges showed significant progress in object picking from a cluttered scene, yet object placing remains challenging. Pose-aware placing based on human and machine readable pieces on an object is useful. For example, the *brandname* of an object placed on a shelf should be facing the human customers. Similarly, the *barcode* of an object placed on a conveyer should be facing a machine scanner. There are robotic vision challenges in the object placing 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 placing system 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 a conveyor with desired poses. Our major contributions include 1) providing an open benchmark dataset of objects, brandnames, and barcodes with multi-view segmentation for training and evaluations; 2) carrying out comprehensive evaluations for our brandname-based fully convolutional network (FCN) that can predict affordance and grasp to achieve pose-aware placing, whose success rates decrease along with clutters; 3) showing that active manipulation with two cooperative manipulators and grippers can effectively handle occlusions 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/**

I. INTRODUCTION

There is a great need of robotic pick-and-place systems in many domains, ranging from warehouse automation, service robots, or grocery stores. Recently more and more robots applied to works in the factory assembly production line or warehouse to reduce human labors. The world-class competitions of Amazon Picking/Robotics Challenges (APC/ARC) 2015-2017 further brought together the teams to develop the picking systems for known or unknown objects from the shelves or totes. The common workflow for 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), estimation of object poses (e.g., geometric model fitting methods such as iterative closest points [5]), selection of objects to pick (e.g, estimate the probability of picking success [6]), and grasping the selected objects. Predicting grasping locations using learning approaches have also been extensively studied [7], [8], [9], [10].

Fig. 1: We utilize product brandname, one of the "semantic labels," for pose-aware placing. We carry out active manipulation with two cooperative robotic arms to handle object occlusions (e.g., brandname facing down). 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 complete placing.

Although picking and grasping prediction have been addressed, *placing* is still not much addressed in previous work, especially when an object needs to be placed not just by its geometry. Such scenarios include product stocking, inventory taking, checkout, and others. When human customers do the shopping in a store or supermarket, they intuitively pay attention to brandname printed on the product, and therefore the products have to be placed on the shelves with the brandname facing outward. It is usually needed for a human cashier to find where the barcode is printed to identify each product while customers checkout. Therefore, estimates of semantics and geometry of object for pose-aware placing are

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

important to create practical values.

We refer a *semantic label* as a piece of surface on an object that provide not just geometry, but also additional information to facilitate manipulations, such as a reference of 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 labels on an object, namely brandname, barcode, one of the six faces of a cubic object, soda can lid, and etc. This work focuses on placing task with brandname semantic labels.

This paper presents an end-to-end pose-aware placing system that utilizes brandname as the reference of object pose and affordance of grippers. We contribute as follows:

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

b) Active Manipulation with Actions of Two Cooperative Arms and Grippers: Other than *passively* using the result of object detection and pose estimation, we present schemes to *actively* manipulate the captured images in order to maximize the visibility of brandnames and individual objects. Our schemes can not only change the camera viewpoint to see the brandname with least occlusion, but also cooperate multiple robotic arms and grippers to manipulate the object, in order to achieve the desired placing.

c) A Benchmark Dataset with Semantic Labels: Despite the recent progress of object segmentation, training a deep convolution neural network usually requires a huge dataset of labeled training data. Although there has been attempts to handle the constraints of novel objects and limited data, the progress is still limited. In order to train the vision algorithms for semantic labels (brandname and barcode), we construct a dataset that include over 8,000 manually-labeled images with brandnames. The dataset consist of 20 objects, and each with brandname and barcode labels. The dataset includes training data of real and virtual environments, a physical benchmark test set for carrying out placing tasks. The datasets are made publicly available [11].

The remainder of the paper is organized as below. Section II describes the related work on recent advances of affordance prediction, as well as active vision and manipulation. We will describe the proposed cooperative dual-arm system in Section III, and how 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 testing set with clutter for evaluations. Section VII provide extensive experiments for the proposed methods on the datasets. Finally, we discuss the future work in Section VIII.

II. RELATED WORK

A. Active Vision and Manipulation

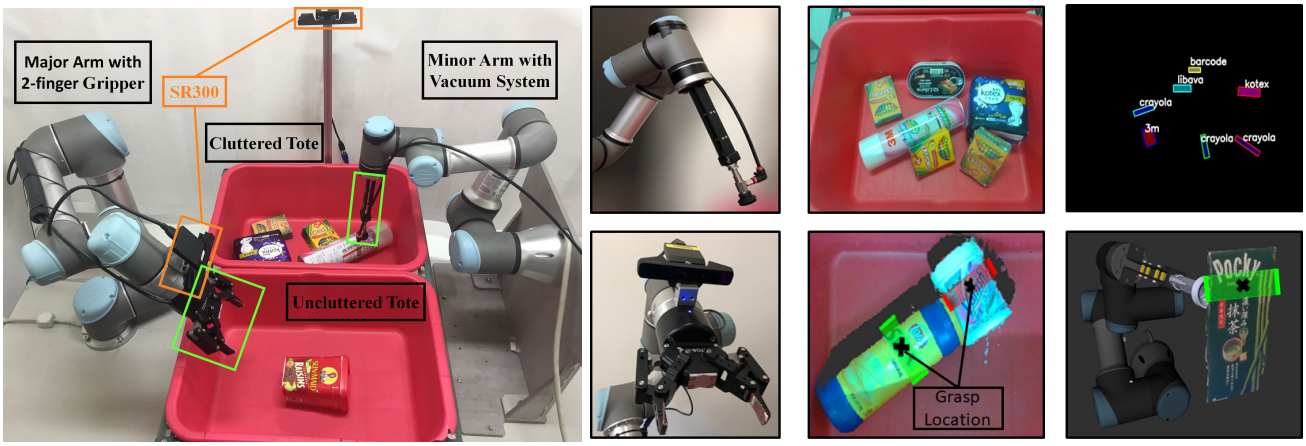
Robotic vision/manipulation, different from the works in computer vision community, has the potentials to control camera or even manipulate with the scene to improve perception [12]. The topics have been adopted in [13] [14] [15] using next-best viewpoint to improve perception confidence for object detection. [6] follows grasp-first-then-recognize strategy to improve poor perception in clutter environments, and further re-orders objects to enable objects being easily grasped by a 2-finger gripper. This line of works fit well to our work that such advantages help overcome the challenge of pick and place in an occluded and clutter environment, and changes the scene to a simpler (uncluttered) environment to obtain higher perception.

B. Affordance and Grasp Predictions

Object affordance (the probability of picking success) is an important topic for pick and place system, and the algorithms tend to be highly related with end-effector co-design. To handle clutter, occlusion conditions, and different object geometry, many recent works adopt different affordance predictions together with a customized end-effector: [16] relies on classic model-based pose-estimation with object registration and decide corresponding affordance modes. [6] defined four primitives for grasping and suction and trained two FCN models to predict the dense pixel-wise affordance probability. Moreover, the affordance, or more precisely grasp prediction for 2-finger gripper to execute picking task in clutter environments. The work in [7] encodes a raw RGB-D image input into several grid cells, and predict direct regression to grasp coordinates under the assumption that there is only a single correct grasp per image. A revised model further predicts multiple grasps per object in real time by using a 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. [17] adopts a multi-stage learning approach which combined CNN and reinforcement learning to learn grasping pose. Subsequently, with two FCNs that map from camera inputs to actions: one for pushes with an end effector orientations and locations, and the other for grasping, the work in [18] used reinforcement learning to decide whether to push/separate adjacent objects or pick objects. All of them show significant progress solving picking problem in clutter environments, but don't consider placements with desired poses.

C. Dual-arm Manipulation

Although manipulation problems have been widely studied with single arm, fewer research works have been investigated in 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 (solves the same task without physical interaction), and bimanual manipulation (physically interact with the same object). [20] demonstrated to execute cluttered picking tasks



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

(b) Brandname-based affordance and grasp predictions.

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

by using dual arm in the goal-coordinated approach: their system only consider how to coordinate in the same working space without interaction. [21] considers object posture and picks and places different shaped object to the box, and may carry out re-grasps object with dual arm in some cases. [22] design a dual-arm system with two different functions: one for sweeping objects with a plate, and the other one for lifting with suction cup in a cluttered container. Our system can be categorized as bimanual, given that the two arms interact while handling a product of brandname facing down. One vacuum gripper arm lifts the object for the other 2-finger gripper arm for picking and placing.

III. System

To execute pose-aware placing under clutter environment, we developed a dual-arm cooperative system which is guided by brandname-based pose-aware affordance prediction, see Fig. 2.

A. System Overview

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

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

3) Vacuum Gripper for Picking and Two-finger Gripper for Placing: To place objects to bin with brandname facing outside, we apply pose-aware affordance to picking objects with 2-finger gripper, see Fig. 2b. However, it's nearly impossible to do in clutter environment due to occlusion. Thus, we adopt 2-stage picking and placing with 2 different-function grippers: vacuum gripper for first picking objects from clutter to uncluttered environment, and 2-finger gripper for second picking objects again and pose-aware placing.

IV. Brandname-based Affordance Prediction

Our dual arm system are guided by the affordance map to decide how to pick object to uncluttered environment with vacuum gripper, and pick and place object to the designed 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 placing. We use brandname as the reference of defining object pose, i.e., the surface normal of the brandname region. We wish to detect a brandname only when its angle between horizontal line ranges from -45 to 45 degrees, for example, not when a brandname is upside down. We train a model of FCN using the training set (further described in Section V), with the model parameters which are initialized from the VGG-DICTNET [23]. It is a convolutional neural network trained from 8 million computer-graphic rendered training data, and is able to recognizes

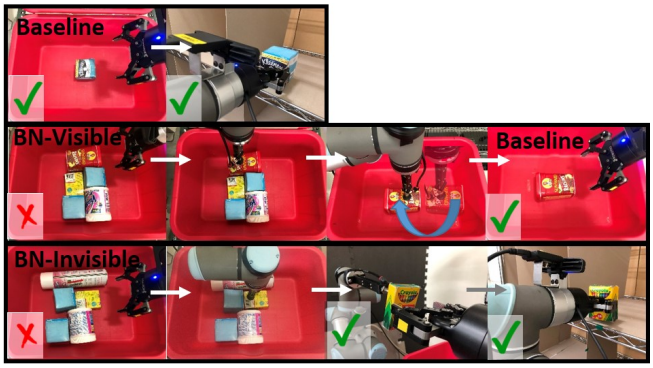


Fig. 3: Action sequences of the baseline (two-finger gripper only) and the proposed active manipulation (cooperative dual-arm system). The action sequences are determined by brandname affordance and grasp prediction that trigger the perception-driven finite state machine.

88,172 dictionary words used for text classification. We modified the fully connected layers into fully convolutional layers using the same schema of the FCN [1], resulting in VGG-DICTNET-FCN. The VGG-DICTNET-FCN model takes a grayscale image as the input and returns a set of 21 densely labeled pixel probability maps, including 20 for the brandnames and one for the background. We then run this model prediction 4 times, rotated at 0, 90, 180, 270 degrees, on a single camera view, and we expect to see the brandname segmentation from only 1 of the 4 predictions. The connected pixels of the segmentations are then clustered as an image mask. Sometimes brandname pixels occur in two of the predictions when their rotation is close to -45 or 45 degrees, and we chose one with a larger mask. Finally, the masks from the prediction results is used for affordance prediction.

2) Affordance Prediction: With the assumption that brandname can be fitted into a rectangle, we calculate an affordance map based on the mask of the segmentation, and further estimate the bounding rectangles to determine brandname pose according to the aspect ratio, in which we assume the brandname has longer width than height. First row of Fig. 2b demonstrates an affordance map predicted from a cluttered tote with 6 objects. For picking with 2-finger gripper, We further estimate grasp locations of two-finger gripper based on the predicted brandname bounding rectangles in affordance map. For picking with vacuum gripper, we search and filter the area in affordance map where surface normal is vertical to the tote and curvature is lower than a threshold as the final picking point.

V. Active Manipulation

In typical pick-and-place case, occlusions and clutters are the challenges to cause failure cases. The proposed brandname-based methods also suffer from self-occlusion or occlusions with the other objects. We introduce our baseline: brandname-based pose-aware placing with 2-finger configuration to show that it could achieve high success rates of pick-and-place in uncluttered scene, but fail if the other object is neighboring the target object in clutters. We further propose a two-stage *active* manipulation system to overcome brandname occlusion, and even the cases while brandname is hidden underneath.

A. Baseline: 2-Finger Gripper Pick and Place

The baseline solution is capable of handling the cases brandname on the object is visible in uncluttered scene: the 2-finger gripper is a good choice to execute pose-aware placing tasks because of the stability. The grasp is estimated using the brandname segmentation pipeline, and we assumed the transforms from brandname pose to grasp pose are known to each specific object and its brandname.

B. Active Manipulation

1) BN Visible: Vacuum Pick-n-place, 2-finger pick-n-place: There are still challenges to successfully pick an object even when the brandname is visible in the cluttered tote, due to the potential collisions with other objects. Previous works in APC/ARC have shown that vacuum gripper outperforms 2-finger gripper during picking stage. Thus, we adopt the strategy of using vacuum gripper to pick an object and place it to the center of the other uncluttered tote, and then conduct another pick and place with 2-finger gripper. 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, 2-finger Pick-n-Place: If brandname is invisible, objects need to be lifted, in order to obtain the brandname underneath. We then carry out the second pick-and-place using a 2-finger gripper to reach a shelf as destination. However, during the first stage we can only rely on object-level FCN. The affordance prediction is based on object segmentations and surface normal of point cloud to determine a picking point, where there should not be object edges or boundaries.

VI. The Brandname Benchmark Dataset

To our knowledge, our Brandname Benchmark Dataset is the first dataset targeting pose-aware placing 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 in 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 groundtruth of objects and brandnames (or other semantic label 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 batch-training (training on object A only or object B only images enables successful prediction on images with both objects A and B) [26], [25]. 3) A test set containing scenes where there are multiple objects in clutter, either duplicated

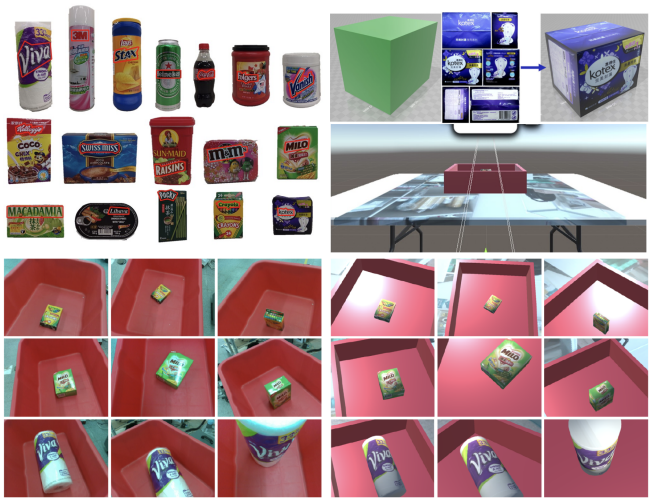


Fig. 4: Data collection of training data in real (left) and virtual (right) environments. Upper-left: Samples of objects and their brandnames (red polygons). Top-right: creating an object model via the 3D builder in Unity. Each scene contains only one object, given the findings that batch-training with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

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 real environments, shown in Fig. 5.

A. Training Sets in Real Environments

The data collection follows previous work [25] in Amazon Picking Challenge 2016. In each scene, single object is placed in the tote with an arbitrary initial pose. A RGB-D camera mounted on the UR5 is used to systematically capture images in multiple *views* automatically. We select 20 products in our dataset, and the selection is based on the following criteria: 1) the physical height of brandname region is at least 1.5 centimeters, 2) single line brandname, and 3) the brandname instance only occurs once on the object. There are 920 scenes $\times$ 31 views. There are 22 missing data from the total 28,520 images due to hardware issues, resulting in 28,498 images.

1) Objects: For the pixel-wise object label generation, we utilize 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 only contain single object and tote in one scene, so there is only tote that would be labeled as background. Due to the monotone of the background, a 2-class output classifier model (2 classes for one is object and the other is background) which 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 with manually labeling a small part of our original images around 500 images using the annotation tool LabelMe [27]. We then fine-tune the VGG16s-FCN network to the 2-class output classifier model using those hand-label images as training dataset to predict the pixel-wise object segmentation of the remaining pictures. Since the prediction result might have scatter noises, we filter the largest area of each object class as final label. Although the prediction results might not be identical to manually labels, the approximations with a certain level of high IoU (intersection over union) are able to allow the robotic arms to compete the tasks.

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

2) Brandnames: In order to train a rotation-variant brandname detector, we expect the predicted brandname is above a certain IoU. Thus, a rotation-variant label criteria is designed for brandname: brandname is labeled only if its angle with horizontal line is from -45 to 45 degrees and at least approximately 50% or more are visible, resulting in about 30% of 28498 images. Such brandname areas are labeled as polygons.

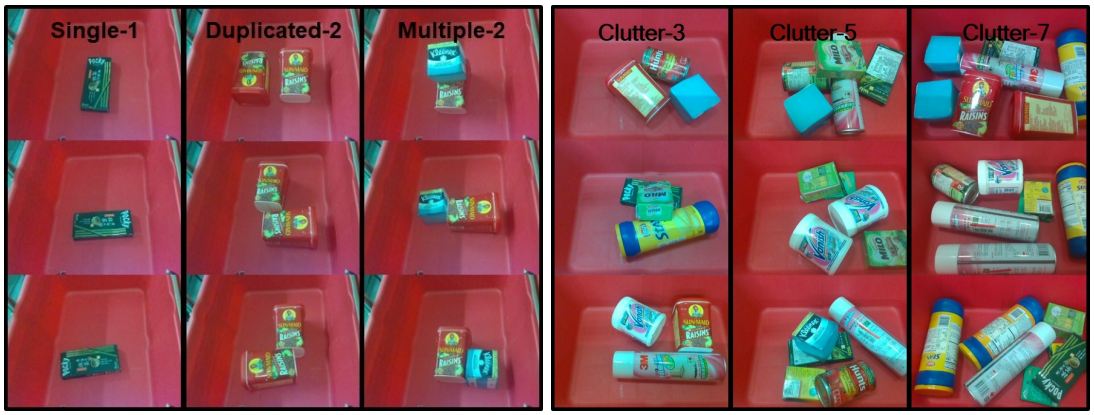
B. Training Sets in Virtual Environments

1) Building Virtual Environment: To make virtual environment as similar as possible to real world, we build a tote with similar configuration in Unity and randomly adjust the hue of light during collecting data, see Fig. 4. For the object model, to avoid distortions and have high resolution model, it is manually created in CAD software 3D Builder [28] and its texture is importing by six faces of high-resolution images from real object. Those CAD models are then labeled with brandname and barcode for each object. Given those settings, we can efficiently collect RGB image, object, brandname, and barcode labels automatically.

2) Data Augumentation: images captured in real environments are often degraded by motion blur and out of focus Gaussian noises. Thus, we include those noises for the collected virtual data to improve the varieties of our virtual dataset. We include two levels of both motion blur and out of focus, resulting in 4 times more data.

C. Physical Benchmark Test Sets

Different from 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 face up in the Single-1, Duplicate-2, and Multiple-2, but some are invisible due to occlusions. There are 20 scenes with various object placements and occlusions among multiple objects in Clutter-3, Clutter-5, and Clutter-7 subsets. In total, there are 290 scenes, 710 objects, and 476 visible brandnames manually annotated for evaluations.



(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.

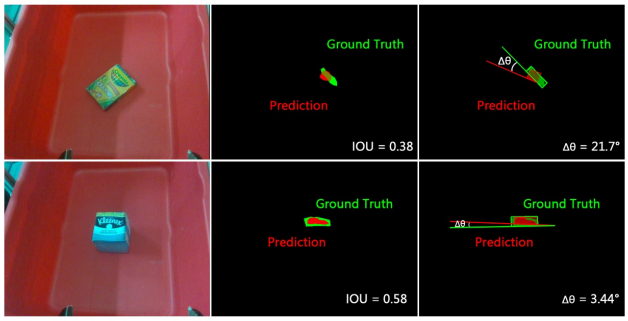


Fig. 6: Samples of IoU and $\Delta\theta$ of brandname segmentations. Higher perceptage of low IoU cases and more $\Delta\theta$ may cause more incorrect affordance grasp predictions to successfully compete pose-aware placing.

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

Benchmark	Image-level		Brandname-level			
	Scene	F-score	Vis. BN	IoU $<$ 0.5	IoU $\geq$ 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

VII. Experiments

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

A. Brandname Segmentations

We first evaluate the rotation-variant predictions of brandname segmentations in the image-level as well as brandname-level, 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 groundtruth rotated rectangles, shown in Fig. 6. Although the findings in [26] suggest that the batch-training with single object generally works well for multi-object scene prediction, we found that the increasing clutters (number of objects in tote) result in lower the image-level F-score. Clutters also cause increasing percentages of low IoU cases and $\Delta\theta$, indicating lower confidences of subsequent affordance and grasp predictions.

B. End-to-end Pose-aware Placing

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

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

We found the grasp predictions of the brandname segmentations could have high picking success rate 0.92, and overall end-to-end placing success rate 0.88 in uncluttered scene. Those results are comparable with the success rate in literatures, such as [9]. Nevertheless, when number of objects increase, the success rates drop, especially in Clutter-3, Clutter-5, and Clutter-7 subsets, where picking with 2-finger gripper does not seem feasible. The baseline solution is not able to deal with BN-DOWN cases with just single arm and gripper.

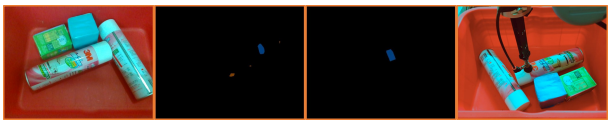
TABLE III: Evaluations of end-to-end placing. As described in Section VI, all brandnames of Single-1, Duplicated-2, and Multiple-3 are facing upward (BN-UP). In other subsets, BN-DOWN represents objects with brandname facing downward. We found that baseline could perform well in unclutter scene but the performance is severely affect by clutter. Active manipulation shows the capbibility of handling BN-DOWN cases, and can retrieve 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

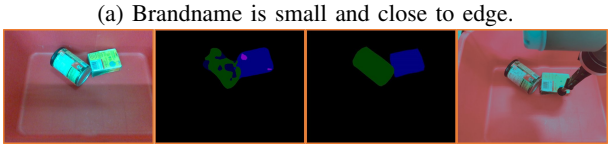
We then evaluate active manipulation, which retrieve objects from clutters in the first picking, and then try the second pick and place. Thus we further use the metrics "First Pick Succ.," if the suction cup stably suck the item without falling. The "Second Pick Succ." follows the baseline "Pick Succ." We found that the success rate of the first pick reaches 0.85 and above in 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, given brandname-visible, it does not guarantee the gripper can grasp the object in clutter environment due to the occlusion. Active manipulation also shows better results than baseline, in which the success rate in placing objects decrease 34%, 23% and 24% in Clutter-3, Clutter-5 and Clutter-7 conditions.

Some common failure cases of vacuum gripper are shown in Fig. 7 and include:

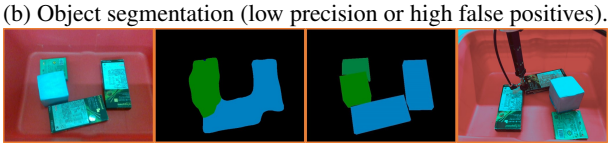
- Vacuum system tends to fail when the affordance prediction is close to the unflattened surfaces of cylinder (Fig. 7a), edge of cuboid objects.
- Poor segmentation results affect the affordance prediction, leading to the slight shifts or rotations of the picking affordance. Such case tends to fail during the placing stage. although the picking stage might work.
- If the duplicated objects are adjacent to each other while brandnames are invisible, it causes trouble 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, making our vacuum system fail to pick them up.



(a) Brandname is small and close to edge.



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



(c) Duplicated objects while brandname facing downward.

Fig. 7: Common failure cases of vacuum gripper. From left to right: input image, brandname or object segmentations, groudtruth labels, and vacuum gripper actions.

VIII. CONCLUSIONS

This work utilizes the properties of one of the semantic labels (i.e., brandname) to predict brandname-based affordance and grasp that complete pose-aware placing tasks in clutter environments. We show that with dual-arm active manipulation enables robots to retrieve information even underneath the occlusions, regardless of the brandname facing upward or downward. Such affordance and grasp predictions are driven by deep models trained in our well-labeled training sets and benchmark testing set. 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, it is worthy of noticing that 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, and may improve affordance and grasp predictions in clutter environments.

ACKNOWLEDGMENTS

The research was supported by Ministry of Science and Technology, Taiwan (grant 107-2923-E009-004-MY3), and Delta Electronics.

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] (2019) 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] (2013) 3d builder. [Online]. Available: https://www.microsoft.com/zh-tw/p/3d-builder/9wzdncrfj3t6?activetab=pivot:overviewtab