IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 March 2023, accepted 27 April 2023, date of publication 5 May 2023, date of current version 11 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3273289

RESEARCH ARTICLE

Multi-Step Object Extraction Planning From Clutter Based on Support Relations

TOMOHIRO MOTODA[1,3], (Member, IEEE), DAMIEN PETIT[1], TAKAO NISHI[1], KAZUYUKI NAGATA[2], WEIWEI WAN[1], (Senior Member, IEEE), AND KENSUKE HARADA[1,3], (Fellow, IEEE)

[1]Graduate School of Engineering Science, Osaka University, Osaka 560-8531, Japan
[2]Future Engineering Research Center, Reitaku University, Chiba 277-8686, Japan
[3]Industrial CPS Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, Japan

Corresponding author: Tomohiro Motoda (tomohiro.motoda@aist.go.jp)

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant JP22J11376, Japan.

ABSTRACT To automate operations in a logistic warehouse, a robot needs to extract items from the clutter on a shelf without collapsing the clutter. To address this problem, this study proposes a multi-step motion planner to stably extract an item by using the support relations of each object included in the clutter. This study primarily focuses on safe extraction, which allows the robot to choose the best next action based on limited observations. By estimating the support relations, we construct a collapse prediction graph to obtain the appropriate order of object extraction. Thus, the target object can be extracted without collapsing the pile. Furthermore, we show that the efficiency of the robot is improved if it uses one of its arms to extract the target object while the other supports a neighboring object. The proposed method is evaluated in real-world experiments on detecting support relations and object extraction tasks. This study makes a significant contribution because the experimental results indicate that the robot can estimate support relations based on collapse predictions and perform safe extraction in real environments. Our multi-step extraction plan ensures both better performance and robustness to achieve safe object extraction tasks from the clutter.

INDEX TERMS Deep learning in grasping and manipulation, logistics, factory automation, manipulation planning, bimanual manipulation.

I. INTRODUCTION

In a logistic warehouse, human workers usually pick and place products from a shelf into a box for service delivery. To replace this logistic operation with a robot, the robot must be able to search for the target product and safely extract it from a shelf in which many products are randomly placed. Thus far, several learning-based methods [1], [2] have designed the motions for robots picking objects from clutter. Picking systems adopted in [3] and [4] used a learning-based grasp detection and action decision model to handle the difficulty involved in picking a specific target from a complex scene. However, extracting the target object from a shelf imposes a new challenge, because a robot needs to safely extract the object while preventing the fall of neighboring objects. To address this problem, our previous method [5] generated a single-step motion plan for selecting and extracting target objects while supporting the surrounding objects. However, it remains difficult for the robot to effectively and safely manipulate products in various scenarios, such as unstable objects, which often require a sequential process to remove objects without disturbing the remaining clutter.

Fig. 1 shows a scenario in which our multi-step motion planner is effective. Here, the robot extracts the target object, box 0, marked in white from the pile. Boxes 1 and 2, however, are stacked on the target. The robot is expected to remove these boxes and subsequently extract the target object. To this end, we need the information regarding where box 2 is supported by box 1 and box 1 by box 0.

In this study, the support relations of the objects included in the clutter are expressed by a graph structure. For example, the support relations of the boxes shown in Fig. 3 can be

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

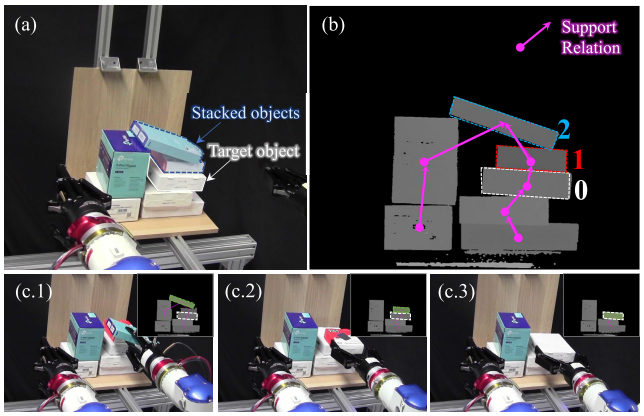


FIGURE 1. Safe object extraction based on support relations. The support relations in the upper right are visualized in a collapse prediction graph. To extract the target object marked in white, the robot extracts the object in a safe extraction order.

expressed by a hierarchically structured graph. To extract box 0 from the clutter, the graph indicates that boxes 2, 1, and 0 should be extracted in this order. This study proposes a novel multi-step object extraction planning from clutter by using graphs obtained by estimating the support relations of objects included in the clutter. Our proposed method analyzes the cluttered environment in the graph structure by considering the physical phenomenon of collapse, and the results of this study are expected to contribute to the development of safe robot manipulation.

The proposed multi-step object extraction planning contains three major components: 1) a collapse predictor (CP) that predicts the support relations between two objects from the clutter by using depth images, 2) a collapse prediction graph (CPG) that consists of the support relations to ensure safe extraction, and 3) a multi-step extraction planner based on the CPG. We infer support relations using a CP based on a deep neural network proposed in [6]. The predictor can predict the movement of objects when extracting an object and identify supported objects for different targeted objects using only depth images. The CPG consists of inferred support relations and provides the best extraction planning by searching for the target object via a recursive traversal search. Additionally, to efficiently extract stacked objects, we propose a novel bimanual extraction planning based on the CPG and validate typical scenes.

The rest of this paper is organized as follows. In Section II, we review related studies. Section III describes the proposed method. In Section IV, we evaluate robotic experiments. In Section V, we discuss the limitations and possible future extension. Finally, we present our conclusions and future work in Section VI.

II. RELATED WORK

Picking objects from clutter is an active research area [1], [7], [8], [9], [10], [11]. This review of related works particularly focuses on three aspects: picking from clutter, visual detection of object relations, and support relations of objects.

A. PICKING FROM CLUTTER

Picking objects is a fundamental task in logistics [3], [12], [13], [14]. Recently, several studies, such as [3], [4], [15], and [16], have assumed that objects are randomly placed in a box and use deep neural networks to perform high-level picking tasks with accurate grasp detection in various scenarios. Mahler et al. built a large-scale dataset, Dex-Net, and predicted the grasp poses based on a convolutional network (ConvNet) [15]. To address a wide range of object categories in cluttered environments, Zeng et al. developed a system for specific target retrieval by pick-and-place with multiple ConvNets [3]. Matsumura et al. adapted the ConvNet to predict object entanglement [16]. Other approaches assumed that objects were placed side by side on a shelf and relocated to pick the target object. Huang et al. planned a sequence of pick-and-place actions to search for an occluded target [17]. Nam et al. proposed relocating objects by pushing [18]. However, if objects are piled on a shelf, an object cannot easily be pushed and slid. By contrast, our study considers the support relationships among stacked objects and presents a multi-step extraction plan to extract the target object. Our method provides a safe extraction process that prevents the fall of neighboring objects.

B. VISUAL DETECTION OF OBJECT RELATIONS

Although the grasp detection algorithms for robotic grasping have achieved significant progress, some of these methods assume the grasp of a single isolated object. In a cluttered environment, a robot should sufficiently understand the cluster to interpret various properties included in an image. These properties include the geometrical, spatial [19], [20], [21], [22], and linguistic [23] relation among objects. Zhang et al. proposed the visual manipulation relationship network to address the grasping order of vertically stacked objects [24], [25] and considered the visual relation of overlap between objects. Recently, datasets such as the visual manipulation relationship dataset [24] and relational grasp dataset [26] have been proposed to build inference models for such relationships. However, these relationships only show the geometrical relationships among objects but cannot be extended to a more general situation of clutter. In this study, the support relations are inferred from a predictive model trained with object movements based on a physics simulation.

C. SUPPORT RELATIONS OF OBJECTS

The support relations among objects have been obtained by analyzing geometric and spatial relationships [27], [28], [29], [30], [31]. Panda et al. extracted the geometrical properties of objects from images and inferred support relations [27], [28]. Kartmann et al. extracted physically plausible support relations using primitive shapes [31]. Paus and Asfour inferred the relations in probabilistic representation, including

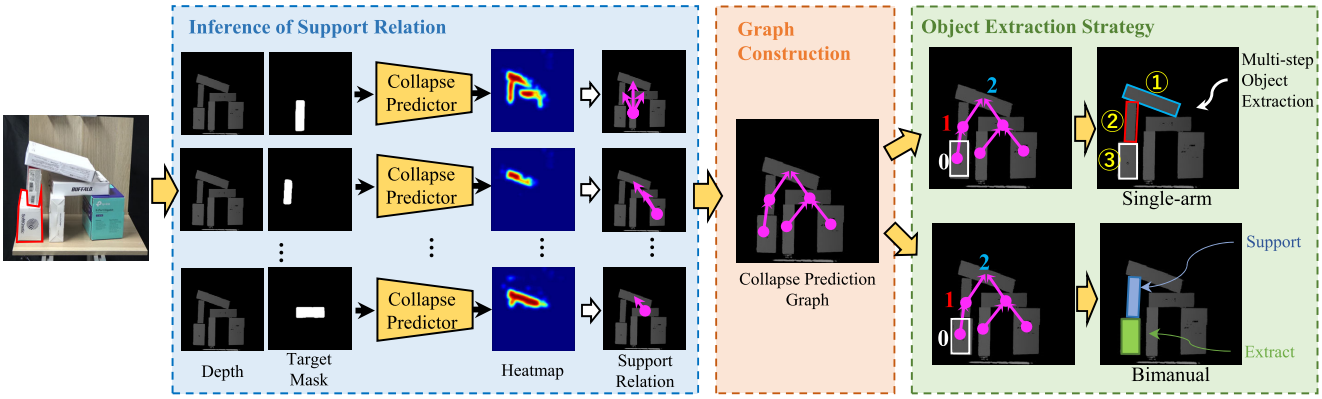


FIGURE 2. Proposed system overview. The collapse predictor outputs the probability that the other objects might fall. Support relations are estimated from this result and graphically represented. Based on the collapse prediction graph, the robot successively picks objects from the pile. In bimanual manipulations, the robot directly extracts the target object by holding the supported object if the target object is supporting only a single supported object.

uncertainty in shapes and poses [32]. In support-relation detection, the related works use the simple primitive that requires preprocessing and pose estimations to approximately predict the scene and restricts real-world use. Contrarily, our study only uses depth images to predict the object collapse and infer the support relations among objects without depending on the shapes and numbers of objects.

III. METHODOLOGY

An overview of the multi-step extraction planning is illustrated in Fig. 2. The proposed framework consists of a CP, the inference of support relations, and a safe extraction strategy. First, we begin with the details of the CP proposed in our previous study [6] (Section III-A). Then, we infer the support relations, which represent the physical relationship between two objects using the CP given a depth image captured from a shelf scene (Section III-B). By concatenating all the support relations, we create a CPG to determine which objects can be extracted from the pile. Herein, we generate a multistep plan to extract the target object (Section III-C). Furthermore, we propose bimanual manipulation based on the CPG for efficiently extracting the target object. The proposed method is described in the following section.

A. COLLAPSE PREDICTOR

The CP is a deep neural network based on the model proposed in [6] and further customized to infer support relations in cluttered environments. This section describes the network architecture, data collection process, and training details. Our method needs sufficient accurate predictions to infer physical relations among objects. Therefore, we extend the dataset and adjust the network parameters to improve the accuracy compared with that of previous studies. The details are as follows.

1) NETWORK ARCHITECTURE

The neural network architecture includes two encoders and a decoder. The scene encoder compresses the input of depth

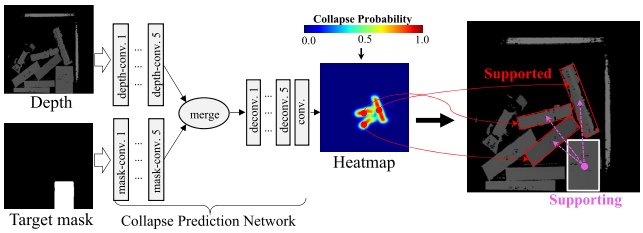
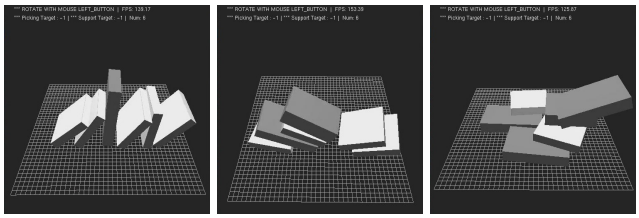


FIGURE 3. Architecture of the collapse predictor consists of two encoders that compress the depth image (256 × 256) and target mask (256 × 256). These outputs are concatenated, and a decoder network generates a heatmap (256 × 256), showing the probability of an object falling. Finally, the support relation is inferred based on the heatmap.

images (256 × 256) with a grayscale using the VGG16 [33] (until the last convolutional block) pre-trained with ImageNet [34]. The first ten convolutional layers are fixed in training to transfer feature extraction. The target encoder converts target masks (256×256) into feature maps using five convolutional (Conv) layers, each followed by batch normalization and rectified linear unit activation layers, respectively. The convolution layers comprise 16, 32, 32, 32, and 64 layers. The Conv layers output latent codes ($8 \times 8 \times 64$). The decoder upsamples the latent code concatenated with both outputs, the head of VGG-16 ($8 \times 8 \times 512$) and target encoder ($8 \times 8 \times 64$), using five Conv layers and one Conv layer. The networks output a heatmap (256×256), which shows the probability of falling objects in pixels. The architecture is shown in Fig. 3.

2) GENERATING TRAINING DATASET

In this section, we introduce the process for our collapse dataset generation. A PhysX physics simulator [35] simulates object removal. First, we place the objects in any of the following scenes: (a) shelved, (b) stacked, and (c) random (see Fig. 4). In the shelved scene, we arrange objects vertically at random intervals, i.e., bookshelves; in stacked, we randomly place objects on each object; and in random, we drop objects at random poses and heights. Specifically, the random scenes



(a) Shelved (b) Stacked (c) Random

FIGURE 4. Simulating the extraction of a box from a clutter. Each scene is generated by adjusting the object poses/positions (a,b) and random pose (c) and dropping the random points (c) on top of others (b) and horizontally (a).

encompass depth data, which we incorporate into the dataset to discern the relative positioning of objects along the z-axis. However, occlusion is beyond the scope of the current study, and therefore shelved and stacked environments were focused on over random environments. In each simulation, we use 5-10 objects in five types of rectangular shapes. Then, a target object is randomly selected and removed from the shelf. During data generation, if the change in the object's center position exceeds a threshold, the objects are moved. We empirically set the threshold to 5.0 mm, coefficient of static friction to 0.9, coefficient of dynamic friction to 0.8, coefficient of restitution to 0.1, and density to 1.0 kg/m^3. Notably, the viewpoint is set to face the shelves, implicitly assuming that the direction of gravity is downward.

We collect a depth image, target mask, and collapse-labeled image. The depth image is a 256 × 256 grayscale height map showing an initial scene in which objects are placed. The target mask is a 256 × 256 binary image in which all the pixels are black, except those of the target object. The collapse-labeled image is also a 256 × 256 binary image in which other objects are annotated after the target is removed.

For data collection, we executed all the simulations in 10,000 shelved, 10,000 stacked, and 30,000 random scenes. The dataset of 50,000 simulations is split into training (90%) and validation (10%). As a test set, we prepared 1,000 simulated data in random scenes.

3) TRAINING DETAILS

The batch size is 24, learning rate is 0.001, and total epoch is 100 with an early stopping with loss monitored. In this study, the training process stopped at 58 epochs. Moreover, the background occupies the heatmap within a wide range, and the network estimates the risk of collapse as lower than the real. Herein, we used the focal loss from RetinaNet [36] as follows:

$$L(y) = -\alpha_y(1-y)^\gamma \log(y), \tag{1}$$

where y is the probability that the predicted labels are equal to the ground truth $\in \{1, 0\}$, $\alpha_y \in [0.0, 1.0]$ is the weight for y, and $\gamma \geq 0$ is the focusing parameter. Intuitively, this scaling factor decreases the contribution of easy examples, i.e., a black background. In our training, α_1 and α_0 are set to 0.25 and 0.75, respectively, and γ is set to 2.0. Table 1 compares the proposed model with that of previous works. The improved model achieves high pixel accuracy, Intersection over Union (IoU), and precision values by using the focal loss. Therefore, we use a weighted model to predict object collapse in later sections. Fig. 5 illustrates the outputs of the trained network.

TABLE 1. Comparison with our previous work.

Model	Pixel Acc.[1]	IoU	Prec.[2]
Previous work	0.984	0.559	0.734
Our method	**0.985**	**0.578**	**0.740**

[1] Pixel accuracy
[2] Precision

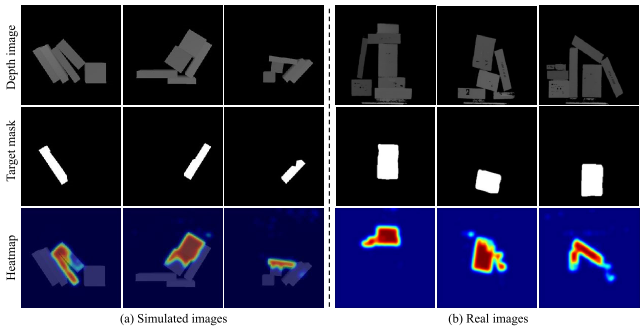


FIGURE 5. Outputs of the collapse predictor. From a set of both the depth images (top row) and target masks (middle row), the proposed network outputs the heatmaps (bottom row), which are the probabilistic color-scale ∈ [0.0, 1.0]. (a) The three images on the left are simulated and (b) the three on the right are real scenes.

B. INFERENCE OF SUPPORT RELATION

In this section, we infer support relations based on the CP. Support relations have been defined in [29], [30], and [31]. Summarily, given two objects X and Y, X supports Y is denoted as SUPP(X, Y). X is the supporting object, and Y is the supported object, i.e., if we remove X from the relation, Y falls. Herein, we focus on the fact that the CP detects objects that fall after removing a target object. Based on this definition, the CP is considered appropriate for detecting the relations between supporting and supported objects.

The flow of inference is as follows. First, we divide point clouds captured with a depth sensor into each object, which provides its target mask and object area R_O. Then, the CP outputs a probability map, which is a dense pixel-wise heatmap with values ranging from 0.0 to 1.0. We calculate the area in the heatmap above the threshold value as the collapse area R_C. If an object is in the collapse area, we consider it a support object for a target, i.e., a supporting object. To detect supported objects, we use the IoU between the collapse area and the area of each object:

$$\text{IoU} = \frac{R_O \cap R_C}{R_O \cup R_C}. \tag{2}$$

This indicates that the overlap ratio of each object with respect to the predicted collapse is R_C. If the IoU exceeds

a certain value, the two objects have a support relation. In a cluttered environment, removing an object may cause several objects that are not in direct contact with the object to fall. When using the CP, such indirect relations between objects should be excluded. Each object is detected as a bounding box (BB), and we evaluate adjacency scores with the IoU based on object BBs that are larger than the original ones. If an adjacency score exceeds the threshold, the relation is considered a pseudo-direct contact.

C. MULTI-STEP OBJECT EXTRACTION

We construct a CPG to determine the next best target that can be safely extracted from the clutter. Given all the support relations, a tree is built with the target object as the root. As shown in Fig. 6(a) and 6(b), we connect the support relations and remove them except for those between adjacent targets.

Our strategy exploits the CPG and safely removes other objects iteratively until the target is extracted. The procedure is shown in Algorithm 1. The multi-step algorithm selects the strategy to safely pick the selected target o_t from all objects O. Initially, we create the CPG G from a depth image Im. In each iteration of the loop (lines 3-21), we extract objects that interfere with the safe picking of the selected target. In line 4, we find the safest object to pick in the clutter. Safe extraction requires selecting a child node for a parent node to minimize the risk of collapse. We explore the CPG by reverse level order traversal with reference to [27] and [28]. If the objects are supported hierarchically, the leaf node, which is not supported by any other object, can be safely extracted in the CPG. Therefore, leaf nodes are extracted first. In a special scenario wherein the parent node has multiple child nodes, we retain a relation between the child and parent nodes at the lowest layer and ignore the other relations (see Fig. 6(c)). This is because if a part of the supported objects is ignored when picking an object at a lower node, a collapse will occur. In lines 4-5, the robot grasps an object o selected from the clutter based on the CPG G.

When pulling an object of a certain width (lines 7-18), we monitor each step of the execution for the potential collapse of objects. Because this research does not consider dynamics during manipulations, an object may fall because of unexpected contact or friction. Therefore, we divide an action into several steps and ensure safety by predicting a collapse score cp before each step. The score is calculated using the collapse area R_C and manipulated object area R_O as follows:

$$\mathrm{cp} = \frac{\mathrm{area}(R_C \cap R_O)}{\mathrm{area}(R_O)}, \tag{3}$$

where area(R) indicates the area of R. cp is the rate of change of R_C of each object, and is used to detect the occurrence of collapse in R_O. In this case, even if only part of the collapse area of the object is detected, the effect of the manipulation cannot be ignored. Therefore, the exact matching of the area is not considered, only the detection rate. If cp exceeds the threshold $\mathrm{cp}_{\max}$, we can re-determine the extraction order to select the other removable object (see lines 8-13 in Algorithm 1). We set the threshold $\mathrm{cp}_{\max}$ to 0.05 considering minor output errors.

If support relations are detected on SUPP(X, Y) and SUPP(Y, X), i.e., supporting each other, we select only the support relationship with the higher collapse score and ignore the other. Then, we determine the extraction order. Notably, when removing these supporting objects with a single arm, bimanual arms should be used.

Bimanual manipulation is relevant for both efficient and safe extractions of clutter. In our previous work [5], we proposed the picking of objects while supporting other objects, but this cannot be applied to robotic picking under limited working ability, such as multiple objects stacked on

Algorithm 1 Multi-step Object Extraction Planning

Input: All objects in clutter O and selected target o_t

1: Im ← Take depth image;
2: G ← Create Collapse Prediction Graph with Im and O;
3: **while** *selected target* o_t *is not extracted* **do**
4: o ← Select the extractable object from G;
5: g ← Generate grasp pose for o;
6: Grasp object o in g;
// Detect the collapse during the object extraction
7: **while** *true* **do**
8: Im ← Take depth image;
9: cp ← Compute collapse score with Im and o
10: **if** cp > $\mathrm{cp}_{\max}$ **then**
11: Release object o;
12: Exit the loop;
13: **end if**
14: Pull object o forward;
15: **if** *object o has been extracted to a certain place* **then**
16: Exit the loop;
17: **end if**
18: **end while**
19: Im ← Take depth image;
20: G ← Renew Collapse Prediction Graph with Im;
21: **end while**
22: **return** Success

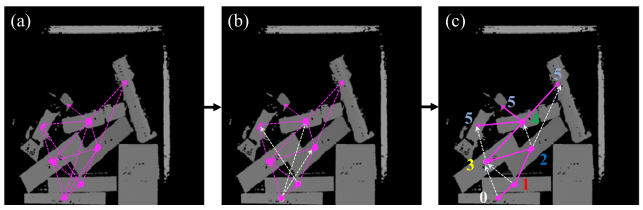


FIGURE 6. Creating a CPG. (a) We connect support relations, create a CPG G on a given object, and (b) remove relations except those between adjacent targets. (c) In these scenarios, the relations connecting to parent nodes at higher child nodes (white edges) are pruned to maintain crucial relations. The numbers indicate the hierarchy based on the target object. The search is conducted via a reverse level order traversal, and each number represents its depth from the root node. Here, objects numbered 5 are leaf nodes, and can be safely extracted.

a single object. In this study, we perform bimanual manipulation using the CPG and find a strategy to pick selected target. This technique can reduce action steps and pick objects efficiently. Through the proposed bimanual manipulation, we verify the capability of the CPG to pick the selected target safely. This procedure is shown in Algorithm 2. First, we ensure that a robot can retrieve a target object while ensuring sufficient support with the other arm. A robot can perform a bimanual action when only one supported object is related to the target (see Fig. 7(d)). One arm grasps the object to prevent it from falling, and the other extracts the target object (see lines 6-9 in Algorithm 2). If two or more supported objects are present (as in Fig. 7(b.1)), before retrieving the target object, the robot extracts the supported objects to possibly satisfy the condition. The CPG for each supported object is constructed, as shown in Fig. 7(b.2). We select and extract the object with the lowest leaf node from the CPGs (see Fig. 7(c)) to satisfy the condition of the bimanual manipulation in a minimum step (see lines 10-13 in Algorithm 2). For example, in Fig. 7(a), at least six objects should be removed based on the CPG. In contrast, when using bimanual manipulation, a robot can extract a target object after removing only one object.

Algorithm 2 Bimanual Object Extraction Planning

Input: All objects in clutter O and selected target o_t

1: **while** *selected target* o_t *is not extracted* **do**
2: Im ← Take depth image;
3: n ← Count the number of objects supporting o_t;
4: ${o_s}^1, {o_s}^2, \ldots, {o_s}^n$ ← Supported objects of o_t;
5: ${G_s}^i$ ← Create the graphs of ${o_s}^i$ with Im and O;
// Extract the target while supporting the object
6: **if** $n = 1$ **then**
7: ${g_s}^1$ ← Generate grasp pose for ${o_s}^1$;
8: g_t ← Generate grasp pose for o_t;
9: Pull the target o_t while supporting ${o_s}^1$;
10: **else**
11: o_l ← Select the lowest leaf node from all graphs;
12: g_t ← Generate grasp pose for o_l;
13: Pick and remove object o_l with grasp pose g_t;
14: **end if**
15: **end while**
16: **return** Success

IV. EXPERIMENTS

In this section, we evaluate the scene analysis from the estimation of support relations and test robotic experiments in a real-world environment.

A. EXTRACTION OF SUPPORT RELATIONS

We evaluated the estimation of the support relations with reference to [31]. The depth images for several real scenes were captured with a YCAM3D-10L (YOODS Co. Ltd., Yamaguchi, Japan) in front of a shelf. We constructed the

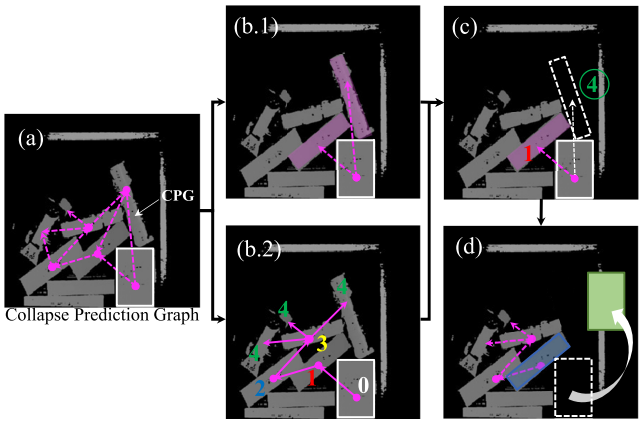


FIGURE 7. Bimanual manipulations based on support relations. Given a CPG (a), we can estimate support relations in contact with the target object marked in white (b.2) and generate the extraction order (b.2). (c) We iteratively remove objects using the extraction order until the target object supports only a single object. (d) The target object marked in green can be safely extracted by fixing the supported object marked in blue.

CPG $G_{\mathrm{HYP}} = (O_{\mathrm{HYP}}, E_{\mathrm{HYP}})$, which is the support hypothesis, using the proposed methods. O denotes objects in the scene, and E denotes a support relation. $G_{\mathrm{GT}} = (O_{\mathrm{GT}}, E_{\mathrm{GT}})$ is generated as the ground truth and manually annotated for the test. In this study, we focus only on the accuracy of the detections of the support relations and ignore the case in which O_{HYP} does not correspond to O_{GT}. Herein, we evaluate our results in terms of precision and recall as follows:

$$\mathrm{Prec} = \frac{|E_{\mathrm{HYP}} \bigcap E_{\mathrm{GT}}|}{|E_{\mathrm{HYP}}|} \tag{4}$$

$$\mathrm{Rec} = \frac{|E_{\mathrm{HYP}} \bigcap E_{\mathrm{GT}}|}{|E_{\mathrm{GT}}|} \tag{5}$$

Table 2 shows precision (Prec.) and recall (Rec.) for 15 scenes and Fig. 8 illustrates selected evaluation scenes. The results of the precision and recall are similar in accuracy to those of the related work [31].

B. REAL-WORLD ROBOT EXPERIMENTS

In all the experiments, we used MOTOMAN-SDA5F (Yaskawa Electric Corporation, Kitakyushu, Japan; a bimanual robot with 15 degrees of freedom (DOFs): seven DOFs in each arm and one DOF in the waist) [37]. The method was programmed using Choreonoid [38] and graspPlugin [39]. The gripper was an adaptive gripper 2F-140 [40] (Robotiq, Lévis, Canada) installed in the arms of MOTOMAN-SDA5F. The YCAM3D-10L was positioned in front of the shelf and could observe the inside [41]. The experimental environment is illustrated in Fig. 9(a). The system used a Core i7-8550U CPU @ 1.80 GHz with 16 G RAM and Python 2.7. The OS was Ubuntu 16.04.

We used 3-10 rectangular objects (see Fig. 9(b)) randomly stacked on the shelf. The robot detected objects by segmenting point clouds with region growing [42] and created a grasp pose from the detected object area.

TABLE 2. Precision and Recall of estimating support relations for all the tested scene images.

Scene	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	Mean
Prec	0.833	1.000	1.000	0.571	0.750	0.750	0.857	1.000	0.750	0.800	1.000	1.000	1.000	0.700	1.000	**0.867**
Rec	1.000	0.667	1.000	1.000	0.750	0.750	1.000	1.000	0.750	1.000	1.000	1.000	1.000	1.000	1.000	**0.928**

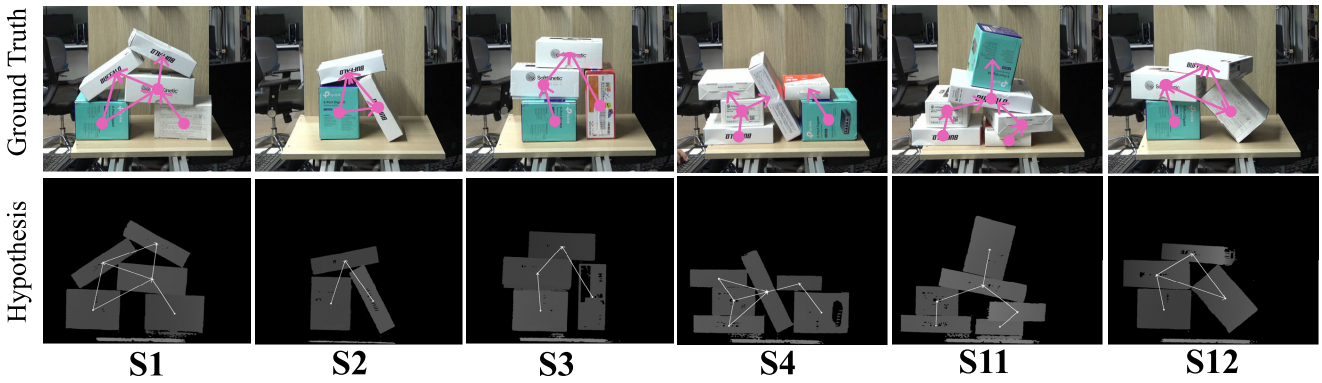


FIGURE 8. Selected evaluation scenes. We estimated the support relations (white edges), except for those between adjacent objects, from the depth images on the bottom row. We sampled the six cluttered scenes on the top row, and manually annotated the support relations (pink edges) as the ground truth.

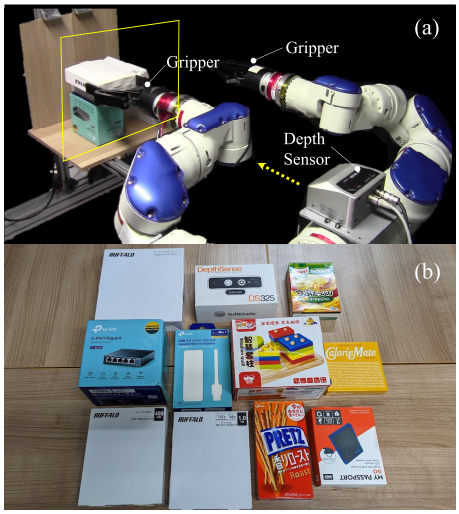


FIGURE 9. (a) Experiment setup, including a MOTOMAN-SDA5F robot, Robotiq 2F-140 grippers, and a YCAM3D-10L depth sensor. (b) Objects used for real-world extraction experiments.

TABLE 3. Real-world extraction performance of different approaches and conditions.

Method	Number of objects	Success rate
Single-step Extraction [5]	3-5	85.0% (51/60)
Single-step Extraction [5]	10	65.0% (13/20)
Multi-step Extraction	**3-10**	**91.2 % (52/57)**

1) EXPERIMENTS ON A SINGLE ARM

These experiments test object picking from a viewpoint whereby support relations are correctly detected. Fig. 10 shows snapshots of the experiments using a real robot; the upper images result from estimating the CPG and extraction order. We conducted 25 experiments using only one-handed manipulation as in proposed Algorithm 1. This algorithm performed well at picking a selected target object with a success rate of 80.0% (20/25), and the mean steps was 2.3.

We compared the proposed method to a single-step method [5]. The single-step method directly extracts the target object based only on initial collapse predictions. The robot attempted to extract a random object from 3–5 or 10 objects using the single-step method and an object from 3-10 objects using the proposed method. The results are shown in Table 3. The success rate at each step was used as the evaluation metric. We achieved the success rate of 91.2% in extracting the objects regardless of the number of objects. The proposed method performed better than our previous work (a success rate of 80.0% in extracting the objects).

2) EXPERIMENTS ON BIMANUAL ARMS

To validate bimanual manipulation, we conducted experiments with the bimanual arms of MOTOMAN-SDA5F. Under the aforementioned condition in the first experiments, we determine an effective option using supporting and extracting actions simultaneously. Fig. 11 shows snapshots of the experiments to trigger the safe extraction order based on the proposed CPG. First, the robot captures the scene and detects support relations. If the target supports only a single supported object, the robot directly extracts the target object while supporting the supported object. If other support relations are identified on the target object, the robot removes an object following the extraction order using the single-arm method.

In Fig. 11, the upper row is a stacked scene experiment and the bottom row is a shelved scene experiment. In the former, three objects are supported by the target object, and

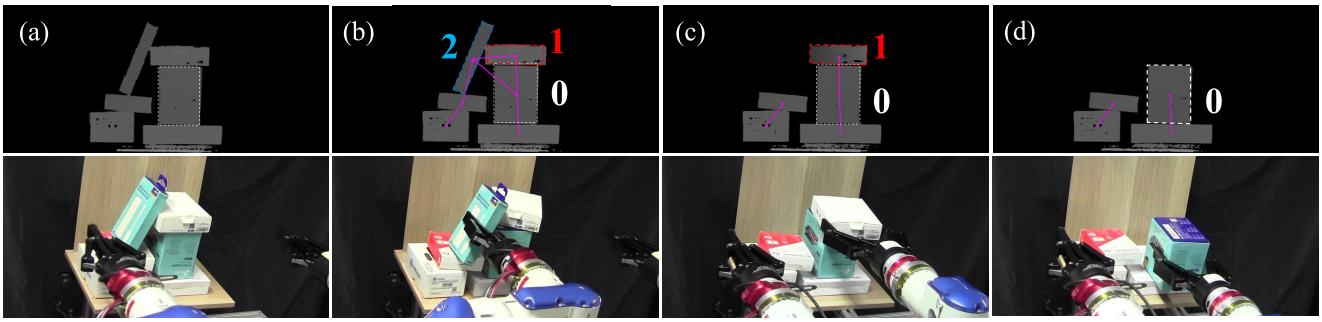


FIGURE 10. Real-world experiment using single arm. (Top) The proposed algorithm estimates support relations from a depth image. A CPG (pink lines) is generated from these relations. (Bottom) The robot selects and extracts an object from the extraction order (a)-(d).

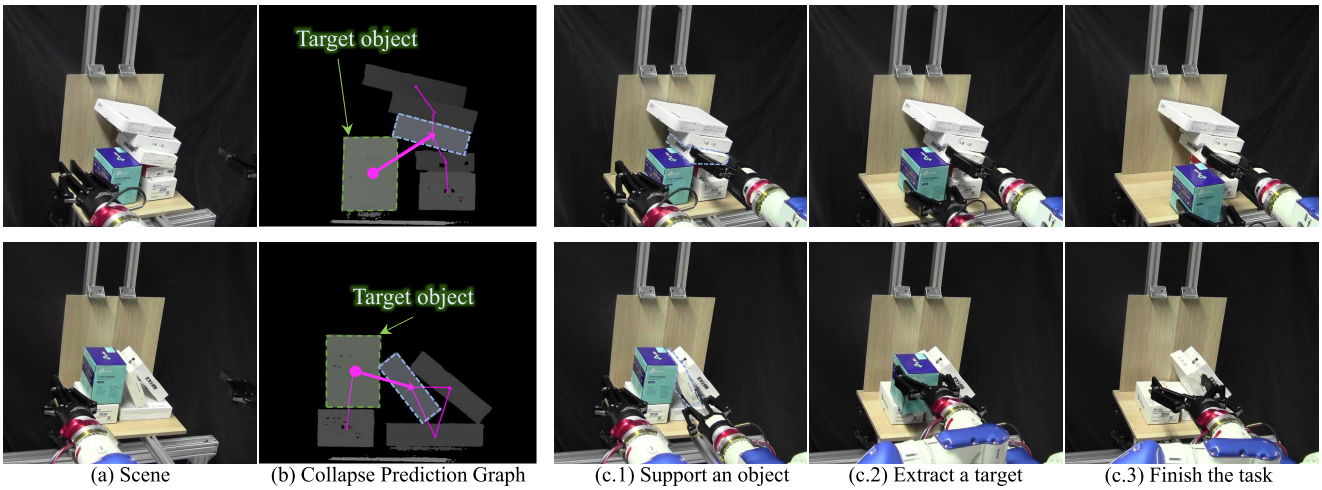


FIGURE 11. Real-world experiments using bimanual manipulations. (a, b) When a target object supports only a single object from the estimated CPG, (c.1) the robot holds the supported object and (c.2, c.3) extracts the target object. (Top) In stacked objects, the robot grasps an object on the target object. (Bottom) When arranging objects horizontally, the robot grasps any object that leans on the target object.

in the latter, two objects are supported by the target object. In each scene, the robot immediately extracts the target object without removing all the supported objects.

V. LIMITATIONS AND POSSIBLE FUTURE EXTENSION

Our study proposes a robotic manipulation system that can safely extract objects from a pile. The experimental results illustrate the importance of identifying support relations and adaptability for safe extraction in the real world. Notably, conventional methods, such as those proposed by [30] and [31], developed the inference of support relations of approximate models from contact and used heuristics with human understanding to predict uncertain information. These studies focused on scene analysis because their applications, which detect a complex scene and real-world manipulation, were problematic. In our study, we proposed a novel multi-step extraction plan and applied it to real-world robotic experiments. Our method achieved more than 90% success in retrieving selected objects by verifying the appropriate extraction order.

Our limitations were observed through physical experiments.

First, learning accuracy has a significant implication for safe manipulation. Missing important support relations can cause damage to the object. As shown in Table 2 and Fig. 9, almost all the support relations are correct. In particular, the recall, which is essential to safely manipulate objects, is higher than 0.9. However, in scenarios similar to S4, where the accuracy is low, the probability of object collapse is often high due to the dense distribution of objects, which result in redundant detection. One of the causes is that internal unobserved parameters such as friction, mass, density, and shape yield unexpected results. To improve detection accuracy, we adjust the trained model on known object shapes [43] and a specific grasp conditions [44]. Moreover, we need inference based on higher-dimensional observation information, such as point clouds, to accurately obtain support relations. Recently, a learning-based model for point clouds has been investigated to extract shape features. Danielczuk et al. proposed a model architecture that examines the collision between point clouds [45]. Chen et al. designed an implicit estimation network to extract a 3D affordance heatmap for each potential task [46]. By using these models, we can accurately detect contact between objects based on

observations. It should also be noted that this study assumes that all objects in clutter are recognizable. Random cases include the problem of detecting occlusions in sensing clutter. In the future, we will extend the inference model to 3D to develop a more robust detection model of support relations.

Second, the arrangement between two arms is challenging. However, detection using the CPG shows that the operation can be performed efficiently with appropriate two-handed manipulation. However, both objects are assumed to be in contact and extremely mutually close. Therefore, the left and right arms can mutually interfere during robot manipulation, and the robot must plan the best motion sequences considering the pose and placement of the two grippers. Objects can still collapse if the supporting hand is removed while another holds the target. In this case, it is more effective to use multi-step extraction with a single arm to provide sufficient safety. Conversely, the purpose of bimanual manipulation is to verify the support relations derived from the CPGs, and developing the algorithm for continuous work is a future task. We must also consider special cases in which two or more objects can support each other, such as SUPP(X, Y) and SUPP(Y, X). In these cases, safe selection is impossible if an operation is limited to using a single arm. This study partially demonstrates that bimanual manipulation can be used in such situations (Section III-C), but as mentioned above, it is an important topic for the future. Recent studies have focused on bimanual manipulation for various tasks. Chen et al. constructed an assembly sequence to evaluate the graspablity, safety, and assemblability of two manipulators [47], [48]. Avigal et al. proposed the BiMaMa-Net architecture for bimanual manipulation, which predicts two corresponding gripper poses without spatial constraint, to improve the bimanual folding for garments [49]. In the future, we will consider using a motion planning method for bimanual manipulation to improve usability.

Thus, identifying support relations from the CP is essential for adaptability to safely pick objects. Our CPG can guarantee a high level of safety in object picking by robots. In the future, we will incorporate lifting and repositioning objects based on action-based physical reasoning. To this end, we will integrate available information, such as the shapes, textures, and masses of objects, to improve the inference model.

VI. CONCLUSION

In this study, we proposed an approach for safe object extraction based on the estimation of support relations between objects. We primarily considered the issue of safe extraction: determining which object should be removed to secure each object from the graph structure by predicting the support relations between supported and supporting objects. This enabled the robot to choose the best next action from the limited observations. Further, a novel bimanual manipulation to directly and efficiently extract the selected target object was proposed. Our proposed method outperformed previous works in terms of success rate, and the performance was improved regardless of the situation, such as the number of objects and arrangements. The experimental results of this study demonstrated that the robot can evaluate support relations by using collapse prediction and perform multi-step safe extractions in real environments, thereby making a significant contribution to the literature.

In future studies, we will learn object movement from time-series data using updated simulations and integrate information on the external properties of objects to predict the outcome of the action. We will also enhance the estimation of support relations by expanding the recognition to 3D space, which is more intricate and practical. To achieve this, we plan to incorporate multi-perspective information into the input of the prediction model.

ACKNOWLEDGMENT

This work was carried out while the author was affiliated with the Graduate School of Engineering Science, Osaka University. The author is now with the Industrial CPS Research Center, National Institute of Advanced Industrial Science and Technology (AIST), and would like to acknowledge the support received from both institutions.

REFERENCES

[1] M. Fujita, Y. Domae, A. Noda, G. A. G. Ricardez, T. Nagatani, A. Zeng, S. Song, A. Rodriguez, A. Causo, I. M. Chen, and T. Ogasawara, "What are the important technologies for bin picking? Technology analysis of robots in competitions based on a set of performance metrics," *Adv. Robot.*, vol. 34, nos. 7–8, pp. 560–574, Dec. 2019.

[2] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, Jun. 2019, Art. no. eaat8414.

[3] A. Zeng et al., "Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 3750–3757.

[4] 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 *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 3347–3354.

[5] T. Motoda, D. Petit, W. Wan, and K. Harada, "Bimanual shelf picking planner based on collapse prediction," in *Proc. IEEE 17th Int. Conf. Autom. Sci. Eng. (CASE)*, Aug. 2021, pp. 510–515.

[6] T. Motoda, D. Petit, T. Nishi, K. Nagata, W. Wan, and K. Harada, "Shelf replenishment based on object arrangement detection and collapse prediction for bimanual manipulation," *Robotics*, vol. 11, no. 5, p. 104, Sep. 2022.

[7] H. Zhu, Y. Y. Kok, A. Causo, K. J. Chee, Y. Zou, S. O. K. Al-Jufry, C. Liang, I.-M. Chen, C. C. Cheah, and K. H. Low, "Strategy-based robotic item picking from shelves," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 2263–2270.

[8] J. Winkler, F. Balint-Benczedi, T. Wiedemeyer, M. Beetz, N. Vaskevicius, C. A. Mueller, T. Fromm, and A. Birk, "Knowledge-enabled robotic agents for shelf replenishment in cluttered retail environments," in *Proc. Int. Conf. Auton. Agents Multiagent Syst. (AAMAS)*, May 2016, pp. 1421–1422.

[9] J. K. Li, D. Hsu, and W. S. Lee, "Act to see and see to act: POMDP planning for objects search in clutter," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 5701–5707.

[10] M. Costanzo, S. Stelter, C. Natale, S. Pirozzi, G. Bartels, A. Maldonado, and M. Beetz, "Manipulation planning and control for shelf replenishment," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 1595–1601, Apr. 2020.

[11] N. Abdo, C. Stachniss, L. Spinello, and W. Burgard, "Robot, organize my shelves! Tidying up objects by predicting user preferences," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 1557–1564.

[12] C. Eppner, S. Höfer, R. Jonschkowski, R. Martín-Martín, A. Sieverling, V. Wall, and O. Brock, "Four aspects of building robotic systems: Lessons from the Amazon picking challenge 2015," *Auto. Robots*, vol. 42, no. 7, pp. 1459–1475, May 2018.

[13] Y.-S. Su, L.-F. Yu, H.-C. Wang, S.-H. Lu, P.-S. Ser, W.-T. Hsu, W.-C. Lai, B. Xie, H.-M. Huang, T.-Y. Lee, and H.-W. Chen, "Pose-aware placement of objects with semantic labels—Brandname-based affordance prediction and cooperative dual-arm active manipulation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 4760–4767.
[14] M. Schwarz, A. Milan, A. S. Periyasamy, and S. Behnke, "RGB-D object detection and semantic segmentation for autonomous manipulation in clutter," *Int. J. Robot. Res.*, vol. 37, nos. 4–5, pp. 437–451, Oct. 2018.
[15] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. Aparicio, and K. Goldberg, "Dex-Net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," in *Proc. Robot., Sci. Syst. XIII*, Jul. 2017, pp. 1–10. [Online]. Available: https://www.roboticsproceedings.org/rss13/p58.html, doi: 10.15607/RSS.2017.XIII.058.
[16] R. Matsumura, Y. Domae, W. Wan, and K. Harada, "Learning based robotic bin-picking for potentially tangled objects," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 7990–7997.
[17] H. Huang, M. Danielczuk, C. M. Kim, L. Fu, Z. Tam, J. Ichnowski, A. Angelova, B. Ichter, and K. Goldberg, "Mechanical search on shelves using a novel 'bluction' tool," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 6158–6164.
[18] C. Nam, J. Lee, S. H. Cheong, B. Y. Cho, and C. Kim, "Fast and resilient manipulation planning for target retrieval in clutter," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 3777–3783.
[19] V. Tchuiev, Y. Miron, and D. Di Castro, "DUQIM-Net: Probabilistic object hierarchy representation for multi-view manipulation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2022, pp. 10470–10477.
[20] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller, "Multi-class segmentation with relative location prior," *Int. J. Comput. Vis.*, vol. 80, pp. 300–316, May 2008.
[21] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.
[22] B. Rosman and S. Ramamoorthy, "Learning spatial relationships between objects," *Int. J. Robot. Res.*, vol. 30, no. 11, pp. 1328–1342, Jul. 2011.
[23] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 852–869.
[24] H. Zhang, X. Lan, X. Zhou, Z. Tian, Y. Zhang, and N. Zheng, "Visual manipulation relationship network for autonomous robotics," in *Proc. IEEE-RAS 18th Int. Conf. Humanoid Robots (Humanoids)*, Nov. 2018, pp. 118–125.
[25] H. Zhang, X. Lan, S. Bai, L. Wan, C. Yang, and N. Zheng, "A multi-task convolutional neural network for autonomous robotic grasping in object stacking scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 6435–6442.
[26] H. Zhang, D. Yang, H. Wang, B. Zhao, X. Lan, J. Ding, and N. Zheng, "REGRAD: A large-scale relational grasp dataset for safe and object-specific robotic grasping in clutter," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 2929–2936, Apr. 2022.
[27] S. Panda, A. H. A. Hafez, and C. V. Jawahar, "Learning support order for manipulation in clutter," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 2013, pp. 809–815.
[28] S. Panda, A. H. A. Hafez, and C. V. Jawahar, "Single and multiple view support order prediction in clutter for manipulation," *J. Intell. Robot. Syst.*, vol. 83, no. 2, pp. 179–203, Mar. 2016.
[29] R. Mojtahedzadeh, A. Bouguerra, E. Schaffernicht, and A. J. Lilienthal, "Support relation analysis and decision making for safe robotic manipulation tasks," *Robot. Auto. Syst.*, vol. 71, pp. 99–117, Sep. 2015.
[30] M. Grotz, D. Sippel, and T. Asfour, "Active vision for extraction of physically plausible support relations," in *Proc. IEEE-RAS 19th Int. Conf. Humanoid Robots (Humanoids)*, Oct. 2019, pp. 439–445.
[31] R. Kartmann, F. Paus, M. Grotz, and T. Asfour, "Extraction of physically plausible support relations to predict and validate manipulation action effects," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 3991–3998, Oct. 2018.
[32] F. Paus and T. Asfour, "Probabilistic representation of objects and their support relations," in *Proc. Int. Symp. Exp. Robot.*, 2021, pp. 510–519.
[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2015, pp. 1–14.
[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 240–255.
[35] *NVIDIA Developer*. Accessed: Oct. 26, 2022. [Online]. Available: https://developer.nvidia.com/physx-sdk
[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.
[37] *Industrial Robots & Robot Automation Tech | Yaskawa Motoman*. Accessed: May 1, 2023. [Online]. Available: https://www.motoman.com/en-us/products/robots/industrial/assembly-handling/sda-series/sda5f
[38] *Choreonoid Official Site*. Accessed: May 1, 2023. [Online]. Available: https://choreonoid.org/en/
[39] *graspPlugin for Choreonoid*. Accessed: Nov. 7, 2022. [Online]. Available: http://www.hlab.sys.es.osaka-u.ac.jp/grasp/en/
[40] *Robotiq: Start Production Faster*. Accessed: May 1, 2023. [Online]. Available: https://robotiq.com
[41] *YOODS*. Accessed: May 1, 2023. [Online]. Available: https://www.yoods.co.jp/products/ycam.html
[42] T. Rabbani, F. A. van den Heuvel, and G. Vosselmann, "Segmentation of point clouds using smoothness constraint," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. 36, no. 5, pp. 248–253, Jan. 2006.
[43] H. Tachikake and W. Watanabe, "A learning-based robotic bin-picking with flexibly customizable grasping conditions," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 9040–9047.
[44] S. Wakabayashi, S. Kitagawa, K. Kawaharazuka, T. Murooka, K. Okada, and M. Inaba, "Grasp pose selection under region constraints for dirty dish grasps based on inference of grasp success probability through self-supervised learning," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 8312–8318.
[45] M. Danielczuk, A. Mousavian, C. Eppner, and D. Fox, "Object rearrangement using learned implicit collision functions," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 6010–6017.
[46] W. Chen, H. Liang, Z. Chen, F. Sun, and J. Zhang, "Learning 6-DoF task-oriented grasp detection via implicit estimation and visual affordance," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2022, pp. 762–769.
[47] H. Chen, W. Wan, and K. Harada, "Planning to build soma blocks using a dual-arm robot," in *Proc. IEEE Int. Conf. Develop. Learn. (ICDL)*, Aug. 2021, pp. 1–7.
[48] H. Chen, W. Wan, K. Koyama, and K. Harada, "Planning to build block structures with unstable intermediate states using two manipulators," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 4, pp. 3777–3793, Oct. 2022.
[49] Y. Avigal, L. Berscheid, T. Asfour, T. Kröger, and K. Goldberg, "SpeedFolding: Learning efficient bimanual folding of garments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2022, pp. 1–8.

TOMOHIRO MOTODA (Member, IEEE) received the B.E., M.E., and Ph.D. degrees from Osaka University, Toyonaka, Japan, in 2018, 2020, and 2023, respectively. In 2022, he was a Research Fellow with the Japan Society for the Promotion of Science (JSPS), where he researched motion planning for robotic picking and replenishment. He is currently a Research Scientist with the Industrial Cyber-Physical Systems Research Center, National Institute of Advanced Industrial Science and Technology (AIST). His research interests include deep learning in grasping and manipulation, motion planning, and manufacturing automation.

DAMIEN PETIT received the M.Eng. degree in robotics and computer vision from Télécom Physique Strasbourg, the M.Sc. degree in robotics and computer vision from the University of Strasbourg, Strasbourg, France, in 2010, and the Ph.D. degree in robotics from the University of Montpellier, Montpellier, France, in 2015. From 2012 to 2015, his research was mainly conducted at the CNRS-AIST Joint Robotics Laboratory, Tsukuba, Japan, and the Interactive Digital Human Group of LIRMM, University of Montpellier, within the European Commission Virtual Embodiment and Robotic Re-Embodiment (VERE). Since 2016, he has been a Researcher with the Graduate School of Engineering Science, Osaka University, Toyonaka, Japan. His current research interests include vision and robot learning, robotic manipulations, and human–robot interactions.

TAKAO NISHI received the Ph.D. degree in agriculture from Okayama University, Japan, in 1999. After working with the National Institute of Advanced Industrial Science and Technology (AIST), he has been an Associate Professor with the Graduate School of Engineering Science, Osaka University, since 2020. His research interests include computer vision, intelligent robotic systems, and agricultural machinery.

KAZUYUKI NAGATA received the B.S. and Ph.D. degrees in engineering from Tohoku University, Japan, in 1986 and 1999, respectively. He joined the Tohoku National Industrial Research Institute (TNIRI), at the former AIST of MITI, Japan, in 1986. He was assigned to the Electrotechnical Laboratory (ETL), in 1991. He was assigned to the Planning Headquarters of AIST in 2001. He is currently a Professor with Reitaku University, Kashiwa, Chiba. His current research interests include the control of robotic manipulations and grasp planning.

WEIWEI WAN (Senior Member, IEEE) received the Ph.D. degree in robotics from the Department of Mechano-Informatics, The University of Tokyo, Tokyo, Japan, in 2013. From 2013 to 2015, he was a Postdoctoral Researcher with Carnegie Mellon University, Pittsburgh, PA, USA, under the support of the Japan Society for the Promotion of Science (JSPS). From 2015 to 2017, he was a tenure-track Research Scientist with the National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan. He is currently an Associate Professor with the Graduate School of Engineering Science, Osaka University, Toyonaka, Japan. His research interests include robotic manipulation and smart manufacturing.

KENSUKE HARADA (Fellow, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in mechanical engineering from Kyoto University, Kyoto, Japan, in 1992, 1994, and 1997, respectively. From 1997 to 2002, he was a Research Associate with Hiroshima University, Hiroshima, Japan. Since 2002, he has been with the National Institute of Advanced Industrial Science and Technology (AIST). From 2005 to 2006, he was a Visiting Scholar with the Department of Computer Science, Stanford University, Stanford, CA, USA, and the Leader of the Manipulation Research Group, AIST, from 2013 to 2015. He is currently a Professor with the Graduate School of Engineering Science, Osaka University, Toyonaka, Japan. His research interests include mechanics and control of robot manipulators and robot hands, biped locomotion, and motion planning of robotic systems.