Prioritized Planning for Target-Oriented Manipulation via Hierarchical Stacking Relationship Prediction

Zewen Wu[1], Jian Tang[1], Xingyu Chen[1], Chengzhong Ma[1], Xuguang Lan[1] and Nanning Zheng[1]

Abstract— In scenarios involving the grasping of multiple targets, the learning of stacking relationships between objects is fundamental for robots to execute safely and efficiently. However, current methods lack subdivision for the hierarchy of stacking relationship types. In scenes where objects are mostly stacked in an orderly manner, they are incapable of performing human-like and high-efficient grasping decisions. This paper proposes a perception-planning method to distinguish different stacking types between objects and generate prioritized manipulation order decisions based on given target designations. We utilize a Hierarchical Stacking Relationship Network (HSRN) to discriminate the hierarchy of stacking and generate a refined Stacking Relationship Tree (SRT) for relationship description. Considering that objects with high stacking stability can be grasped together if necessary, we introduce an elaborate decision-making planner based on the Partially Observable Markov Decision Process (POMDP), which leverages observations and generates the least grasp-consuming decision chain with robustness and is suitable for simultaneously specifying multiple targets. To verify our work, we set the scene to the dining table and augment the REGRAD dataset with a set of common tableware models for network training. Experiments show that our method effectively generates grasping decisions that conform to human requirements, and improves the implementation efficiency compared with existing methods on the basis of guaranteeing the success rate.

I. INTRODUCTION

Robot grasping in stacking scenarios has always been a challenging problem. When performing manipulation tasks, robots are required to execute efficiently and ensure the environments are controllable and accident-proof. Therefore, it is indispensable to learn the stacking relationships between objects [1] [2]. Considering a kind of multi-object stacking scenario, where objects are not only in a stacked state but the object below provides full support for the object above, and the latter reaches a stable equilibrium state. This situation widely exists in fields of dish serving, kitchen tidying, logistics transporting, and so on. It is a special case of clutter [3], while its unique attribute contributes to the high speed and efficiency of tasks on the premise that the robot fully understands the environment.

Existing algorithms in learning stacking relationships for manipulation focus on distinguishing simple relationship types. Zhang et al. [4] establish Visual Manipulation Relationship Network (VMRN) to predict the manipulation

[1]Z. Wu, J. Tang, X. Chen, C. Ma, X. Lan, and N. Zheng are with National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, National Engineering Research Center for Visual Information and Application, Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, No.28 Xianning Road, Xi'an, Shaanxi, China `wuzewenchn@foxmail.com`, `xglan@mail.xjtu.edu.cn`
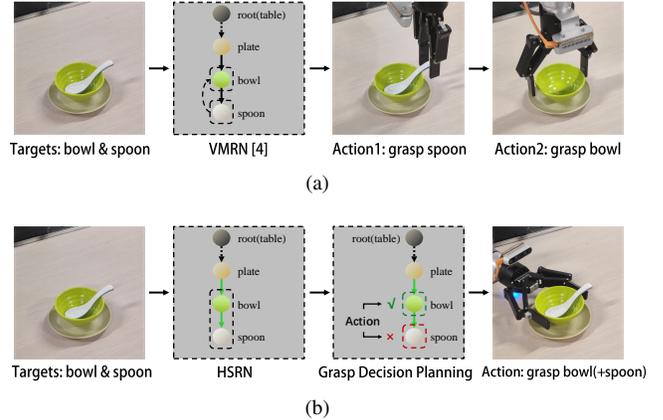
Fig. 1. Grasping in the orderly-stacked scene. The staking relationships are presented by a tree structure. When the spoon is stably supported by the bowl and both are the targets, grasping the bowl directly will be more in line with human common sense and improve execution efficiency. (*a*) Existing method [4] generally perform them separately. (*b*) Our perception-planning method optimizes the grasping decision, only needs to grasp the bowl to complete the task.

relationship of each object-pair in the scene, combine them as a tree structure for presenting results so that a grasping order can be determined when any target is specified, as shown in Fig. 1(*a*). The relevant works [5] [6] optimize the manifestation and performance of relationship detection, but lack further subdivision for the hierarchy of stacking relationship types. In scenarios where objects are stably stacked, it is insufficient for the robot to perform high-efficient grasping.

Inspired by human common sense, our work follows a principle: when we want to grasp a pile of objects that are stably stacked together, we can directly grasp the bottom object, as illustrated in Fig. 1(*b*). For example, when assigning a robot to grasp and deliver a pile of dishes and bowls, it is irrational to perform them separately, but grasp once and for all. This condition will become more intricate when one of the dishes is not required. Another case is when there are multiple sundries stably stacked above the target object, removing the sundries all at once will greatly save time and the number of manipulations. Therefore, we try to solve the problem that how to decide the best grasping action in the current orderly stacking scenario when targets are designated.

In this paper, we present a perception-planning architecture for generating optimal grasping decisions when specifying any number of objects in the scene as targets. We introduce a Hierarchical Stacking Relationship Network (HSRN) for scene perception, which takes dining scene RGB images

as input, and outputs the hierarchical stacking relationship between each pair of objects through object detection and relationship prediction, presented by the Stacking Relationship Tree (SRT). Then a planning algorithm designed from the Partially Observable Markov Decision Process (POMDP) is implemented to comprehensively consider the inadequacy of the perceptual process and the priority of the grasping targets, obtain an underlying estimate of the real state and provide an optimal grasping decision chain. In order to train our network, we augment a virtual orderly-stacked scene dataset on the basis of REGRAD [7], attempting to simulate the stacking environment of dining tables. Experimental results substantiate the feasibility of our perception and planning process, which outperforms state-of-the-art baselines in target-oriented task manipulation while saving considerable time. Our main contributions are summarized as follows:

- A hierarchical stacking relationship prediction network that distinguishes different degrees of stacking between objects.
- A POMDP-based planner for planning optimal manipulation decision chain in the current scene when targets are specified.
- The experimental results confirm that our method improves the efficiency of performing grasping tasks, and is more in line with human habits.

II. RELATED WORK

A. Visual Stacking Relationship for Manipulation

The visual relationship has been studied in the field of computer vision for a long time. In analyzing the relationship between the foreground in the image, the classification consideration is frequently based on the triplets of <subject, predicate, object> as a whole [8] [9]. Other works cope with this problem by predicting subject, object, and their relationship separately [10] [11]. However, this kind of relationship analysis lacks direct guidance for robotic manipulation.

In the scene of stacking cluttered objects, the visual relationship between objects is mainly represented as support and occlusion. Panda et al. [12] define a variety of support relationships between objects, including support from below, support from the side, and containment, constructing the inferred support sequence of objects. Zhang et al. [4] build a Visual Manipulation Relationship Network (VMRN) for representing the stacking relationship in the scene, and collect a dataset for relationship detection in robotic grasping. Yang et al. [13] introduce fully connected Conditional Random Fields (CRFs) on [4], removing redundant relationship representations. Zuo et al. [14] and Ding et al. [5] deploy graph neural networks to collect contextual information about objects and improve relationship detection performance. Tchuiev et al. [6] utilize Deformable DETR [15] as the backbone, representing object hierarchy in directed graph adjacency matrix form. Current approaches focus on pursuing high accuracy of detection, but for orderly-stacked scenarios, a lack of in-depth understanding of the relationship types cannot satisfy diversified task requirements. In comparison,

our work unveils subtle classification and verification of the stacking relationship, which is more in line with human cognitive habits for manipulation.

B. Task Planning with POMDP

Perceived incompleteness emphasizes the necessity of decision-making for efficient completion of manipulation within a task-specific framework. It is not only reflected in the acquisition of complete information, but also in the intensive analysis of task requirements. Such problems can be modeled as POMDP-based approaches. Different from motion planning that considers the spatial movement of the manipulator [16] [17], this type of task planning instructs robot the manipulated objects and methods. It is a high-level manipulation indicator and vitally dependent on task settings [18]. Pajarinen et al. [19] design to grasp objects which may be occluded with special attributes and the occlusion information is estimated for planning the best action to be performed. Li et al. [20] direct at searching for objects in a refrigerator, plan a sequence of actions to rearrange objects, and find the target. Xiao et al. [21] consider object search for fully occluded objects, use parameterized action to deploy manipulation. Recent works [22] [23] introduce human-robot interaction to facilitate the disambiguation of task instructions. Since the human-specified target is probably unknown or the description language is too vague to confirm the target, POMDP is needed to comprehensively consider observations and human commands to plan action of grasping or asking. Inspired by the above works, we formulate the task as human-like grasping on a dining table, aiming to execute actions according to target requirements without redundancy.

III. OVERVIEW

Our perception-planning architecture is shown in Fig. 2, including object detection Ω, relationship prediction Ψ, and decision-making planning process Φ. The model incorporates POMDP planning on the deep learning training algorithm. In the training network, we mainly consider object properties and the hierarchical stacking relationship of each object pair. Our model takes RGB images as input, filters a series of proposals for category recognition and bounding-box regression. According to the detected objects, relationship predictor Ψ then constructs the full permutation of all binary object groups to classify their stacking parent-child relationships which are distinguished by stable support and weak support.

The training results are limited by noise, object occlusion, and dataset size. Thus in order to realize robust grasping execution of the robot on the observation results obtained by scene state, we introduce the POMDP-based planning model. POMDP groups targets designated by humans. For a single target group, it assigns action values to all potential actions and updates the belief, which includes historical optimal state value, providing criteria for the selection of each decision-making step. During the planning process, we chiefly consider whether to grasp certain objects together to target area or non-target area, recursively obtain all possible
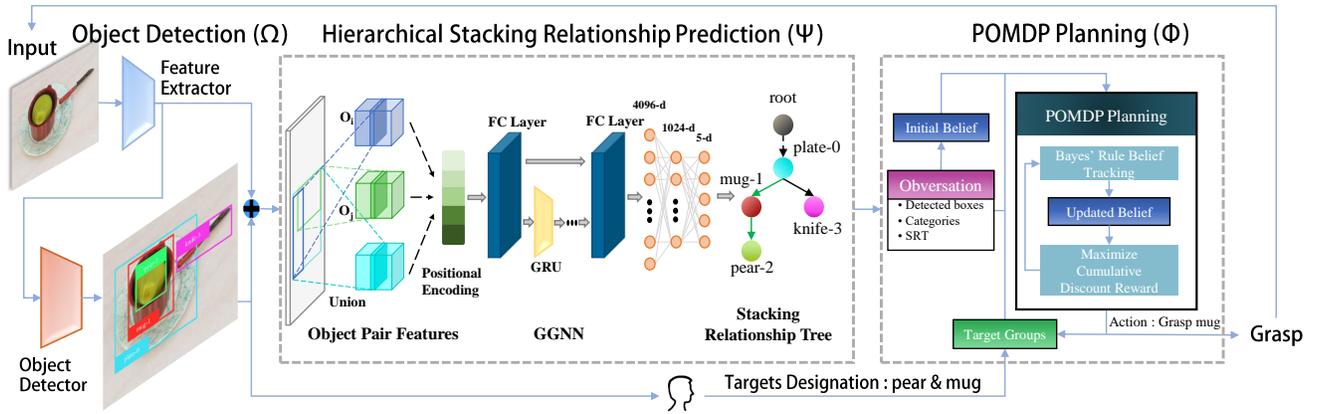
Fig. 2. The overview of our proposed approach. The RGB image first needs to go through feature extraction and object detection, then the relationships between object pairs are predicted through the HSRN. Finally, these perception results will be used as observations, and input into the POMDP planner together with human target designation to plan the current optimal grasping action.

task execution trajectories and at last search out the optimal current grasping action with the highest expected cumulative reward.

IV. PROPOSED APPROACH

A. Feature Extraction and Object Detection

The RGB image I captured by the camera first needs to go through a series of convolutional layers for feature extraction, as the input to generate the candidates for target positioning and recognition. Our work continues the target detection proposals in [4], using ResNet101 [24] as the feature extractor and Faster R-CNN [25] as the object detector. The results of feature, object categories, and bounding boxes will be sent to the relationship prediction network for further analysis.

B. Stacking Relationship Hierarchy Extension

The stacking of objects normally hinders the safety and feasibility of grasping, but some stable stacking can also improve grasping efficiency. Zhang et al. [4] define three stacking relationships between objects. If moving object o_a will change the spatial state and stability of object o_b, i.e. o_a supports o_b, we call o_a is the parent of o_b, on the contrary, o_a is called the child of o_b. This kind of relationship classification is not enough to cover the physical spatial relationship and is not adequate to describe scenes with stable stacking. On this basis, we refer to the definition of the support relationship in [12] and give:

- Stable Support: the parent fully supports the child, when the parent is grasped and generates translation in 3d space without rotation, the child will be grasped as well.
- Weak Support: the parent partially supports the child, and the child is simultaneously supported by other objects, including the operating table.

Fig. 3 presents some cases of Stable Support and Weak Support. In Stable Support relationships, the parents generally have container attributes, like bowls. However, in another critical situation although the parent fully supports the child, due to the limits of material rigidity, friction coefficient,
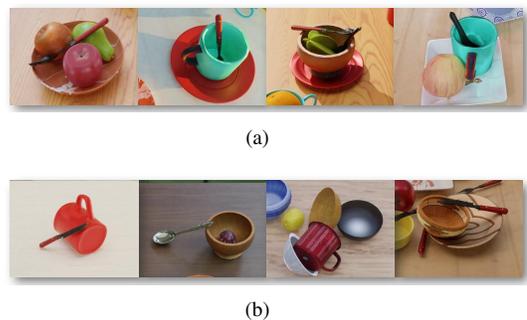

(a)

(b)

Fig. 3. Different hierarchies of stacking. (a) Stable Support. (b) Weak Support.

and robot manipulation ability, grasping the parent with its child is still difficult, e.g. a book with a pen on it. Thus in the decision-making process, we put forward a thorough planning scheme to decide whether the parent object and child object should be grasped together.

C. Relationship Prediction

To predict the relationship properties of each object, our relationship prediction network Ψ extracts the individual and union features of all pairwise combinations of objects $\{(o_i, o_j) | i, j \in N_{obj}\}$ detected by object detector as nodes and apply the Gated Graph Neural Network (GGNN) [5] for detecting relationships. We embed the positional encoding into each object-pair node feature, integrate all of the feature node information through Gate Recurrent Unit (GRU) in a full connection manner, and update the hidden-layer vector h_i^t describing node information. We concatenate the GRU output features with initial hidden-layer features to avoid forgetting. Finally, three linear layers are applied to classify relationships, followed by refined SRT representing the prediction results as shown in Fig. 4(b). The edges between parent and child nodes are distinguished by stable and weak. In order to standardize, we add a root node to join all the relationship trees, which can be regarded as the operation
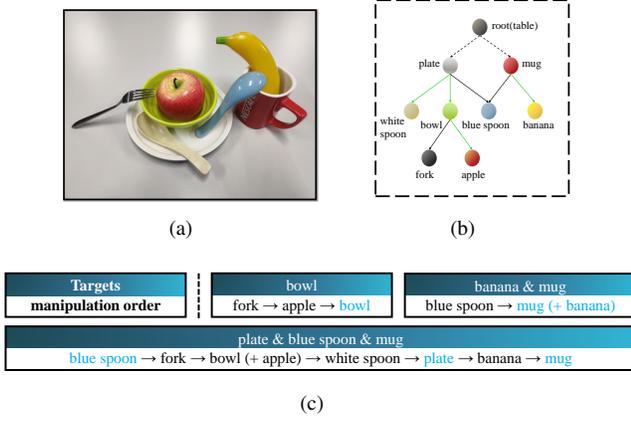
Fig. 4. Relationship prediction and planning results according to specified targets. (*a*) The input RGB image. (*b*) Stacking Relationship Tree (SRT) to present relationship prediction. The green edge represents the parent node and the child node are stacked in Stable Support, and the black edge represents Weak Support. (*c*) The optimal manipulation order chain under the specified targets. The object in parentheses means it will be grasped together. The object in black means it needs to be grasped to the non-target area, and the object in blue is going to be grasped to the target area.

table. According to the description in Section IV-B, classified relationship types R_c can be expressed as follows:

- o_a is ordinary parent (*op*) of o_b (Weak Support)
- o_a is ordinary child (*oc*) of o_b (Weak Support)
- o_a is natural parent (*np*) of o_b (Stable Support)
- o_a is natural child (*nc*) of o_b (Stable Support)
- o_a and o_b have no support relationship

The network uses a multi-class cross-entropy function as the loss function for stacking relationship prediction:

$$L_{RP}(R; \Psi) = -\frac{1}{N_{obj}(N_{obj} - 1)}$$
$$\sum_{(i,j) \in N_{obj}^2, i \neq j} \sum_{c=1}^{5} r_c \log\left(p\left(r_c | o_i, o_j; \Psi\right)\right) \tag{1}$$

and the total loss of our complete network is:

$$L(I; \Omega, \Psi) = \mu L_{OD}(O; \Omega) + (1 - \mu) L_{RP}(R; \Psi) \tag{2}$$

where $L_{OD}(O; \Omega)$ is loss function of object detector Ω, as discussed in [26]. We set the balance weight μ to 0.5, considering the trade-off with two network modules.

D. POMDP-based Planning

A POMDP is modeled as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{Z}, \mathcal{T}, \mathcal{O}, \mathcal{R})$, where $\mathcal{S}$, $\mathcal{A}$, $\mathcal{Z}$ respectively denote the set of states, actions, and observations. Transition function $\mathcal{T}(s, a, s') = p(s' \mid s, a)$ indicates taking action $a \in \mathcal{A}$, the probability from state $s \in \mathcal{S}$ to the next state $s' \in \mathcal{S}$. Observation function $\mathcal{O}(s', a, z) = p(z \mid s', a)$ is the probability of observing $z \in \mathcal{Z}$ by performing action a and reach the resulting state s'. $R(s, a)$ represents the immediate reward of action a in state a. What follows are detailed descriptions of each part.

1) State: We factorize the scene state $\mathcal{S}$ to the state of each object $s_i \in \mathcal{S}$ [27]. s_i is represented by the Boolean quantity s_i^g of whether it is present in the image (or is grasped and removed) and the relationship between it and other surrounding objects $\{s_{ij}^r | j = 1, 2, ..., N_{obj}\}$, the state quantity of o_i is $N_{obj} + 1$. The relationship state space can be provided from R_c, attached to a value indicating the nonexistence of s_{ii}^r, or when o_i is taken away from scenes. Since the real state is unknown due to the partial observation and noise, for each state $s \in \mathcal{S}$, we maintain a belief $b \in \mathcal{B}$ as a state observer to represent the distribution estimation of state. For each object o_i, all its beliefs can be expressed as:

$$\mathcal{B}_i = \{b_i^g\} \cup \{b_{ij}^r | j = 1, 2, ..., N\} \tag{3}$$

where b_i^g refers to the probability that o_i is observed in the image, and b_{ij}^r refers to the observation relationship between o_i and other objects. We update b in real time after each step of decision-making, re-evaluating the state distribution of the current environment.

2) Action: The task in our work is target-oriented, which means human designates the required targets according to the result of object detection, then send it to POMDP for decision-making. As taking grasping an object an action, the action space is $\{2N_{obj} + 1\}$, including grasping each object to target area or non-target area, alongside a report action that no longer exists grasping action to perform.

3) Observation Model: We take the results of object detector Ω and stacking relationship predictor Ψ in HSRN as observations, indicating as Z^g and Z^r. The observation function $\mathcal{O}(s', a, z)$ includes the probability that the updated scene state s' can be accurately observed after performing the grasping action a, expresses the difference between the real state and the robot perception of the scene. Therefore we learn the probability distribution of observation from the average recall rate index of each object category $rc_i \in RC$ of both network modules, as it approximates $p(z_i \mid s_i)$ while o_i is presented in the scene. Since an action a potentially not only causes one object to be grasped, the observation function is comprehensively denoted as $\prod_{j \in C(o_i)} rc_j$, where $C(o_i)$ is the set of o_i and its child nodes.

4) Transition Model: Ideally, a grasping action will lead to a state update where the object is no longer presented and its relationship with other objects becomes non-existent. Nevertheless, the robot's manipulative capacity results in different performances in grasping various types of objects. Therefore we collect a series of empirical data in our robot experimental environment about the capacity to grasp types of objects separately. We choose 15 common item categories in the dining table scenes, for each category, measure the success rate that it can be grasped by the gripper horizontally and stably by performing grasping 20 times. The grasping method [28] is adopted for our data collection. We quantify the average success rate with data normalization to obtain the transition probability $p(s' \mid s, a)$ of each object (see Fig. 5). If grasp fails, we default the state does not change to simplify decision-making reasoning. In experiments, we revise the
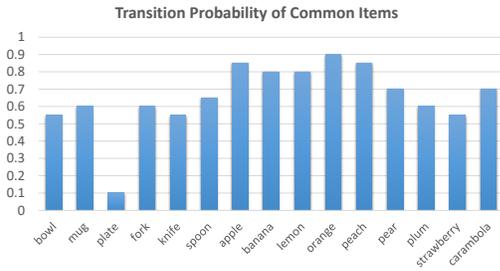
Fig. 5. Transition probability of common object categories.

data based on the real grasping success rate and verify the validity of the transition modeling.

5) Reward Function: We encourage the robot to complete the task with the least number of grasps, so we give the agent a base penalty of -10 for each grasp action. The relationship between the executed object and the other will generate additional rewards, we empirically design a relation-based grasping reward for each object:

$$R\left(s, a_i\right) = -10 + \begin{cases} 5\tanh\left(N_{nc}\left(o_i\right)\right), & \text{if } N_{nc}\left(o_i\right) > 0 \\ -10\tanh\left(N_{oc}\left(o_i\right)\right), & \text{if } N_{oc}\left(o_i\right) > 0 \\ -2, & \text{if } N_{np}\left(o_i\right) > 0 \\ 0, & \text{else} \end{cases} \quad (4)$$

where N_{nc}, N_{oc}, and N_{np} describe the number of natural children, ordinary children, and natural parents of each object. In an SRT, if o_i has natural children, we expect it to be grasped actively and give it a major reward. Hyperbolic tangent function $\tanh\left(x\right) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ fixes reward upper limit. When o_i has ordinary children, it is unwarrantable to be grasped by giving a large penalty. If o_i has a natural parent, it is in principle not encouraged to be grasped directly, with a slight penalty of -2. If o_i is a leaf or isolated node, no reward will be attached. Since we set the scene state value when the task is completed to 0, the purpose of this design is to avoid the optimal state value obtained at each step being greater than 0.

6) Belief Updating & Planning: In partially observable domains, after taking action a, the observation z is received and belief b will be updated to b'. As mentioned in (3), we refine the scene state to the private state of each object and update each belief individually. The planner starts with an initial belief b_0, applies Bayes' rule to track the belief at each step. Specifically:

$$b'\left(s'\right) = \lambda \mathcal{O}\left(s', a, z\right) \sum_s \mathcal{T}\left(s, a, s'\right) b(s) \quad (5)$$

where $\lambda = \frac{1}{p(o|a,b)}$ is a Bayes' rule's constant for normalization.

For the SRT generated by Ψ, we cannot directly reference it for planning, after all our purpose is not just cleaning the table (of course it also can be), but to grasp the desired targets. To this end, we first construct the Descendant Hash Table (DHT) of each object according to SRT. In the table, for each object o_i, all objects in the subtree with o_i as the root are arranged in order from leaf to root, the last element is o_i itself, i.e. the action space when o_i is the target.

It has been reflected in section IV-B that parent-child nodes connected by Stable Support are not necessarily grasped together. After identifying the corresponding prediction results and obtaining the targets designation D, the planning process first needs to group the targets, because stacking relationships may exist between the targets themselves. Our planner incorporates the target nodes connected by the Stable Support relationship into a target group $g \in G$. Since targets and non-targets should not be grasped together, it downgrades Stable Support between target groups and non-targets to Weak. Then the planner sort all groups from leaf to root, sequentially search all descendants of each group based on DHT as action, and implement planning. After all nodes in the current action space are pruned, switch to the next group till the end. In planning process, a policy $\pi : \mathcal{B} \to \mathcal{A}$ is learned to maximize the cumulative discount reward according to the initial belief b_0:

$$V^\pi\left(b_0\right) = \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t \mathcal{R}\left(s_t, a_t\right) \middle| a_t = \pi\left(b_t\right)\right] \quad (6)$$

where the discount factor γ is set to 0.8. Last, carry out the look-ahead search to find the optimal action:

$$a^* = \arg\max_a V^\pi\left(b\right) \quad (7)$$

We concatenate the grasp action of each step in the planning process into a chain as a manipulation decision chain, as presented in Fig. 4(*c*).

V. Experiment

A. Dataset Construction

On the basis of REGRAD [7], we expand the dining scene dataset REGRAD-v2 in a virtual environment, replenish 15 object categories of main tableware objects (plate, bowl, mug, spoon, fork, knife) and some fruits (apple, banana, etc.). The models with an amount of 44 come from Shapenet [29], and YCB dataset [30], all dimensions are based on real objects. In order to prevent model penetration due to the initial z-axis height limit, we use Poisson disk sampling [31] to uniformly select the plane position of model loading, load other tableware and fruits after the container objects are placed. Each scene contains 8 to 12 objects. In addition, we also generate substantial cluttered scenes in REGRAD standard procedure to supplement negative samples. Our newly generated dataset contains 3.2 k scenes and entirely shares all properties and functions with the original dataset. On the basis of the automatic label generation method [7] of object bounding box, object category, 2D grasping position, and manipulation relationship (not distinguish Stable and Weak), we manually label all Stable Support relationships. Some scenes in REGRAD-v2 are shown in Fig. 6.

B. Perception and Decision

1) Implementation Details: Our model is implemented in the PyTorch framework and uses an NVIDIA RTX 3090 with 24GB of memory to train, the maximum training epoch is 30. The object detector used is Faster R-CNN [25] with the
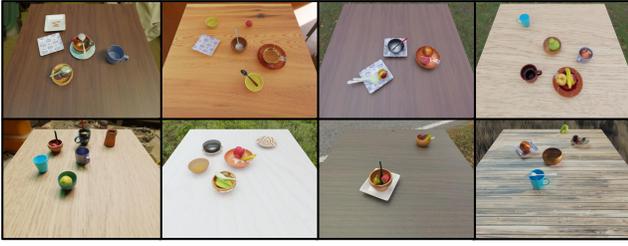
Fig. 6. Some examples of REGRAD-v2. Objects are mostly stacked in order, fully simulating the stacking method on real dining tables.

Algorithm	Metrics (%)			
	mAP	OR	OP	IA
VMRN[4]	90.28	66.78	65.82	21.40
GVMRN-RF[14]	89.90	68.59	68.31	24.06
GGNN-VMRN[5]	90.54	72.26	71.95	26.62
HSRN(ours)	**90.69**	**72.84**	**73.53**	**26.85**

Algorithm	*Task ST*		*Task MT*		*Task TC*	
	AR_f	AR_w	AR_f	AR_w	AR_f	AR_w
VMRN[4]	97.65	96.54	62.40	47.36	76.42	24.88
HSRN	92.30	87.55	75.02	58.59	96.46	**90.81**
HSRN-POMDP	**98.64**	**97.37**	**98.25**	**95.63**	**97.31**	86.52

backbone of pre-trained ResNet101 [24]. We set the learning rate to 0.001 and decayed to 0.0001 after 10k iterations. Stochastic Gradient Descent(SGD) with a momentum of 0.9 is used as the optimizer. All hyperparameters are shared during training.

For each scene in REGRAD-v2, we collect 2 groups of artificially designated single target and 2 groups of multiple targets as task requirements, with an extra task of table cleaning in which all objects are regarded as targets. We specifically annotate the desired grasping decision chain by human anticipation according to the principle in section I. and utilize perception results for planning, in order to verify the effectiveness and rationality of our POMDP model. In the generated chain, for some unrelated objects, the order of grasping does not affect the rationality of the overall decision-making. So we are concerned about whether the robot plans redundant or omits necessary grasping actions rather than pursuing complete alignment with annotated decision chain.

2) Evaluation Metrics: The metrics of the perception process continue the baseline of previous work for visual manipulation relationship detection [4]. On this basis, we design the measurement standard of the planning process.

- Object detection: **mAP** (main Average Precision) measures the average precision of all object categories, which is commonly used in object recognition.
- Relationship prediction: The metrics used to evaluate predicted relationship indicated as triplet $< o_i, r_{ij}, o_j >$ are **OR** (Object triplet Recall), **OP** (Object triplet Precision) and **IA** (Image-Wise Accuracy). **IA** calculates the proportion of all objects and triple relationships are accurately predicted in an image.
- Decision Making: Planning rationality criterion is discussed in Section V-B.1. We refer to the set of all objects in a decision chain as Action Object Set (AOS). In particular, if AOS generated by the planning process is not equal to the AOS of annotated chain, i.e. the grasping actions are redundant or missing, the decision chain is irrational. Since we update the scene and re-decision after each action to ensure robustness, we evaluate the average rationality of the first step decision AR_f and the whole decision chain AR_w separately.

3) Performance: Our perception-planning results are shown in Fig. 7. We compare our method with state-of-

the-art stacking relationship detection baselines performed on the REGRAD-v2 dataset. Implementation settings are the same as our method, using ResNet101 for feature extraction and Faster R-CNN for object detection. Relationship prediction in these baselines only considers original parent-child relationships defined by [4]. Table I shows the results that our algorithm is better than other baselines generally. Although the hierarchy of relationships is more complex in our work, the dining table scenes have more restrictions on the representation of object stacking than clutter scenes, which facilitates our network to learn the staking modes of objects. Furthermore, GGNN integrates context information, fully utilizes global scene description to comprehend Stable and Weak Support, ensuring comprehensive perception performance.

The implementation of POMDP planning will reduce unnecessary grasping steps and make robotic action more in line with human habits. Since existing baselines [4] [14] [5] specify stacking relationship as the manipulation relationship, when the relationship is predicted, they can directly search action decision. The search method is all the same with these baselines, which is subsequent traversing from SRT. Thus we only compare the average rationality AR_f and AR_w of decision-making with [4] baseline, as shown in Table II. Task *ST*, *MT*, *TC* respectively means single target grasping task, multiple targets grasping task, and table cleaning task.

It is clear to see that due to the lack of hierarchical distinction between the relationship types, [4] can only grasp objects one by one, which is irrational especially in task *MT* and *TC* on account of planning redundant grasps. We also give the decision results that do not go through the
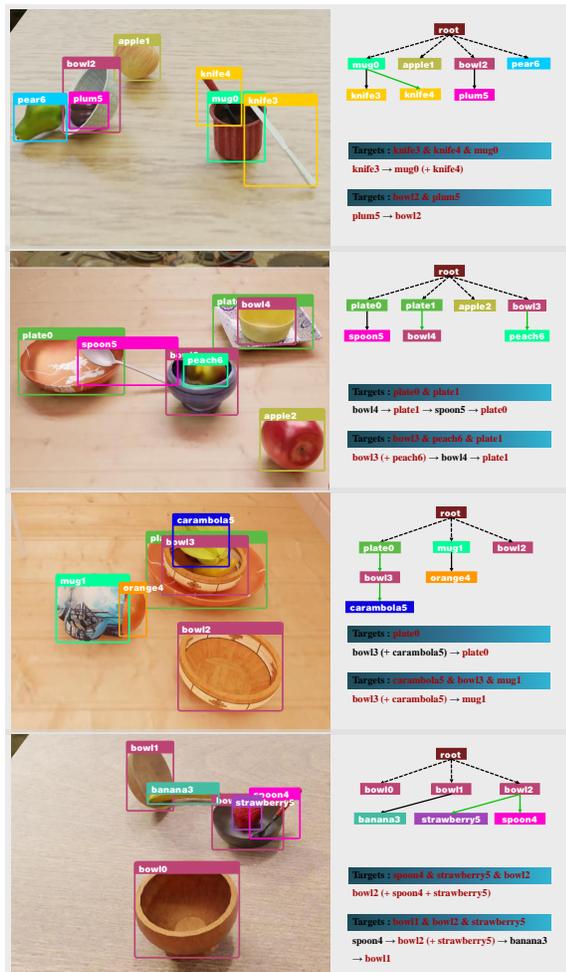
Fig. 7. Perception and planning results of our method. **Left** Object detection results. **Top right** Stacking Relationship prediction results that presented by SRT. **Bottom right** Planning grasp-decision chain based on human-given target designations.

POMDP planning, only perform action relying on HSRN and the principle in Section I that if nodes are connected with Stable Support, grasp the parent node. Apparently, it is arbitrary to plan in this way, amounts of necessary grasps are ignored. It is likely to perform the action which grasping targets with non-targets into the target area together. With the implementation of the POMDP decision-making process, the planning chain with high rationality can be obtained for all three tasks. The POMDP also reflects some compromises made in the design of the robot's capabilities which is unfortunately against human hope. For example, the difficulty of stably grasping the plate is too high, so when the plate has a natural child such as an apple, they will still be grasped separately. So in task *TC*, AR_w of HSRN-POMDP is not as high as directly plan from HSRN. This is the result of considering comprehensive factors.

C. Robot Manipulation

1) Experimental Setup: We use a UR5e robot to perform our grasping experiments. The robot is equipped with a 2-finger Robotiq 2F-140 gripper and an eye-in-hand Intel

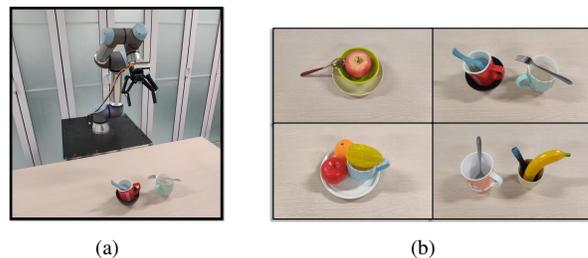| Task	Algorithm	Success Rate (%)	Number of Grasps	Time Cost (min)
ST	GGNN-VMRN[5]	10/12	2.83	2.63
	HSRN-POMDP	**11/12**	**2.08**	**1.89**
MT	GGNN-VMRN[5]	7/12	4.67	4.24
	HSRN-POMDP	**10/12**	**3.52**	**3.20**
TC	GGNN-VMRN[5]	4/12	5.66	5.17
	HSRN-POMDP	**7/12**	**3.58**	**3.32**

(a)   (b)

Fig. 8. Robot manipulation configuration. (*a*) Experimental environment. (*b*) Examples of setting scenes for grasping tasks.

RealSense D435i camera as shown in Fig. 8(*a*). We use ROS to drive the robot and gripper. The real models used include bowls, mugs, tableware, and fruits, that are stacked on the operating table in an orderly manner, as shown in Fig. 8. After the robot perceives the scene and obtains the detection result, it informs humans of the detected object categories with serial numbers to ensure the uniqueness of each object label. The human then specifies the desired targets and the robot plans to grasp them to complete the task.

We set up 12 scenes in total, with 3∼6 objects in each scene, and each scene is divided into three tasks *ST*, *MT*, and *TC*. In experiments, we restore the initial scene as consistently as possible. The criterion for task success is that all specified objects are grasped and placed in the target area. We compare our method with the baseline of state-of-the-art VMRN-based method [5]. We adopt ROI-GD [28] method for our grasp position detection, which is as same as used in Section IV-D.4. We train the detection network based on the 2D grasping position generated during dataset construction. In the experiment, we fine-tune the grasping position by discretized searching based on the detected grasping position, select the position whose projection in the image mostly overlaps the detected bounding box. When grasping objects such as bowls and mugs, our two-finger gripper first adjusts to horizontal orientation, and approaches the object in a horizontal manner. When grasping objects such as fruits, the gripper approaches the object in a top-down direction to conform to human grasping habits.

2) Results: Table III summarizes the success rate, average number of performed grasps, and average time cost of our method with baseline under different task settings. The

results reveal that our network generalizes well to real grasping scenarios and performs better than the GGNN-VMRN method in terms of task completion. In experiments, our method saves an average of 29.48% on task execution time, because some grasping actions are legitimately saved while ensuring execution stability, and the entire manipulation process is more in line with human behavior habits.

Due to the limited width of the jaws, when a plate is in targets, it is barely able to be successfully grasped. Comfortingly, POMDP effectively judges the children stacked on plates as a separate grasp, not limited to Stable and Weak relationship predictions. Other error cases show the incompleteness of decision-making. For example, the fork is stably supported by the mug, but it is in a state of lying across the rim of the mug. Our planning process has a considerable probability of directly grasping the mug. Ideally, there would be no problem with this action, but in practice, this may cause the fork to drop during movement.

VI. CONCLUSION

This paper proposes a hierarchical object stacking relationship detection network and introduces a POMDP-based decision-making process for giving a rational grasping execution process for specific tasks. Experiments show that our algorithm greatly simplifies the grasping process while ensuring the success rate. Future work will pay more attention to the in-depth understanding of the scene and human requirements, analyzing and reconstructing the scene to complete tasks in a more humanized way. Besides, we will pursue further improvement in the detection efficiency of different types of stacking relationships, and provide incisive robotic understanding for robust grasping in multiple scenarios.

REFERENCES

[1] M.-Y. Liu, O. Tuzel, A. Veeraraghavan, Y. Taguchi, T. K. Marks, and R. Chellappa, "Fast object localization and pose estimation in heavy clutter for robotic bin picking," *The International Journal of Robotics Research*, vol. 31, no. 8, pp. 951–973, 2012.

[2] A. Murali, A. Mousavian, C. Eppner, C. Paxton, and D. Fox, "6-dof grasping for target-driven object manipulation in clutter," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6232–6238.

[3] D. Fischinger, M. Vincze, and Y. Jiang, "Learning grasps for unknown objects in cluttered scenes," in *2013 IEEE international conference on robotics and automation*. IEEE, 2013, pp. 609–616.

[4] H. Zhang, X. Lan, X. Zhou, Z. Tian, Y. Zhang, and N. Zheng, "Visual manipulation relationship network for autonomous robotics," in *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2018, pp. 118–125.

[5] M. Ding, Y. Liu, C. Yang, and X. Lan, "Visual manipulation relationship detection based on gated graph neural network for robotic grasping," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 1404–1410.

[6] V. Tchuiev, Y. Miron, and D. Di Castro, "Duqim-net: Probabilistic object hierarchy representation for multi-view manipulation," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 10 470–10 477.

[7] 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 Robotics and Automation Letters*, vol. 7, no. 2, pp. 2929–2936, 2022.

[8] A. Farhadi and A. Sadeghi, "Recognition using visual phrases," in *Computer Vision and Pattern Recognition (CVPR)*, 2011.

[9] S. K. Divvala, A. Farhadi, and C. Guestrin, "Learning everything about anything: Webly-supervised visual concept learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3270–3277.

[10] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 852–869.

[11] K. Liang, Y. Guo, H. Chang, and X. Chen, "Visual relationship detection with deep structural ranking," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[12] S. Panda, A. A. Hafez, and C. Jawahar, "Learning support order for manipulation in clutter," in *2013 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2013, pp. 809–815.

[13] C. Yang, X. Lan, H. Zhang, X. Zhou, and N. Zheng, "Visual manipulation relationship detection with fully connected crfs for autonomous robotic grasp," in *2018 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2018, pp. 393–400.

[14] G. Zuo, J. Tong, H. Liu, W. Chen, and J. Li, "Graph-based visual manipulation relationship reasoning network for robotic grasping," *Frontiers in Neurorobotics*, vol. 15, p. 719731, 2021.

[15] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[16] S.-K. Kim and M. Likhachev, "Planning for grasp selection of partially occluded objects," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 3971–3978.

[17] N. P. Garg, D. Hsu, and W. S. Lee, "Learning to grasp under uncertainty using pomdps," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 2751–2757.

[18] M. Lauri, D. Hsu, and J. Pajarinen, "Partially observable markov decision processes in robotics: A survey," *IEEE Transactions on Robotics*, 2022.

[19] J. Pajarinen and V. Kyrki, "Robotic manipulation of multiple objects as a pomdp," *Artificial Intelligence*, vol. 247, pp. 213–228, 2017.

[20] J. K. Li, D. Hsu, and W. S. Lee, "Act to see and see to act: Pomdp planning for objects search in clutter," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 5701–5707.

[21] Y. Xiao, S. Katt, A. ten Pas, S. Chen, and C. Amato, "Online planning for target object search in clutter under partial observability," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8241–8247.

[22] H. Zhang, Y. Lu, C. Yu, D. Hsu, X. La, and N. Zheng, "Invigorate: Interactive visual grounding and grasping in clutter," *arXiv preprint arXiv:2108.11092*, 2021.

[23] Y. Yang, X. Lou, and C. Choi, "Interactive robotic grasping with attribute-guided disambiguation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8914–8920.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[26] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[27] C. Diuk, A. Cohen, and M. L. Littman, "An object-oriented representation for efficient reinforcement learning," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 240–247.

[28] H. Zhang, X. Lan, S. Bai, X. Zhou, Z. Tian, and N. Zheng, "Roi-based robotic grasp detection for object overlapping scenes," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 4768–4775.

[29] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[30] B. Calli, A. Singh, J. Bruce, A. Walsman, K. Konolige, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Yale-cmu-berkeley dataset for robotic manipulation research," *The International Journal of Robotics Research*, vol. 36, no. 3, pp. 261–268, 2017.

[31] R. Bridson, "Fast poisson disk sampling in arbitrary dimensions." *SIGGRAPH sketches*, vol. 10, no. 1, p. 1, 2007.