2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

RelationGrasp: Object-Oriented Prompt Learning for Simultaneously Grasp Detection and Manipulation Relationship in Open Vocabulary

Songting Liu, Tat Joo Teo, Zhiping Lin and Haiyue Zhu†

***Abstract*— Autonomous robotic grasping under complex, clustered, and unstructured environments is a fundamental but challenging task. To achieve human-like rationality in dealing with the grasping task, the agent requires hybrid intelligence from multilateral aspects. This paper introduces *RelationGrasp*, a unified framework employing a transformer encoder-decoder structure to simultaneously achieve open-vocabulary object detection, manipulation relationship inference, and grasp pose detection. A unique object-oriented prompt learning mechanism is designed to seamlessly bridge the grasp pose and manipulation relationship branches, delivering high fidelity of object-grasp affiliation for object-aware grasping and grasp sequence planning. By formulating the relationship detection as an adjacency matrix regression task under multi-task learning, our framework significantly increases the relationship accuracy with reduced computational overhead. Moreover, to facilitate the robust and adaptive deployment of the proposed *RelationGrasp* to novel environments, we propose a consistency-based self-supervised adaptation strategy to adapt the pre-trained network to new scenarios and improve grasp accuracy on unseen objects. Our proposed network achieved state-of-the-art performance on various public dataset such as VMRD, OCID, etc., in both grasp detection and manipulation relationship classification, and real-world robot experiments has also been conducted to show the practical usages.**

I. Introduction

Robotic grasping is a fundamental but important task for many industrial applications and daily services, which is targeted to enable robots to intelligently interact with and manipulate objects in unstructured environments. Its key objective is to autonomously plan the robot grasping based on the sensor input, i.e., the problem of grasp detection for objects in a given scene. With the recent advancements in deep learning and computer vision, the learning-based grasping approaches have revolutionized grasp detection, achieving remarkable success in many scenarios. State-of-the-art approaches [1]–[3] have demonstrated high accuracy and robustness in the grasp pose inference for both 2D and 3D environments.

Nevertheless, existing solutions still suffer from the high complexity in various real-world scenarios, where multiple objects with diversity can be randomly mixed, clustered, and stacked. Although many works can already achieve the grasp detection for unknown/unseen objects, compared with human intelligence, it still lacks the reasoning capability on high-level planning to determine the proper operation order. In other words, most works only focus on the grasping stability of the targeted object but without taking account of its effects on the surrounding objects. To achieve autonomous grasping for such complex scenarios, not only the grasp detection but also the sequence planning are essential to handle the objects with minimal cross interference. However, most existing research still treat these as two separate problems, i.e., the grasp pose detection and the relationship inference. The former is only responsible to detect all the possible grasp candidates for a given scene, and the latter merely inferences the spatial relationship between objects to identify the manipulation order, e.g., which object is on the top and should be manipulated first, etc.

As most grasp detection approaches are instance-agnostic while the relationship detection are correlated with the specific instance, there lack an intrinsic bridge between the grasp pose and the relationship, and thus raise a challenge to associate the manipulation order with the grasp pose to form the integrated solution. Although some post-processing operations based on intersections of bounding boxes or segmentation masks can be proposed to align the grasps detection with object relationship, e.g., 1) link the detected grasps with objects and 2) map with the relationship through the objects, such framework stacked with multiple networks and post-processing generally leads to large computational overhead. From the deployment aspect, it is desired to have an end-to-end framework that can address the integrated planning (relationship and order) and execution (grasp pose) simultaneously to achieve the autonomous grasping.

In this work, we propose a unified framework, *RelationGrasp*, to achieve the simultaneous prediction of the grasp poses, object relationships, and object detection. Our framework employs the DETR-like transformer encoder-decoder structure, which contains three parallel-yet-interacted decoder branches for respective tasks. We achieve the open-vocabulary object detection by integrating the Contrastive

This research is supported by A*STAR "Control Policy Training for Multi-Robot Collaborative Manipulation with Imitation (C221518002)", and also supported by A*STAR "RIE2025 IAF-PP Advanced ROS2-native Platform Technologies for Cross sectorial Robotics Adoption (M21K1a0104)" programme.

S. Liu is with Singapore Institute of Manufacturing Technology (SIMTech), Agency for Science,Technology and Research (A*STAR), and also with School of Electrical & Electronic Engineering, Nanyang Technological University (NTU), Singapore 639798, (e-mail: lius0081@e.ntu.edu.sg; ezplin@ntu.edu.sg).

T. J. Teo is with the Robotics, Automation & Unmanned Systems Centre of Expertise, Home Team Science and Technology Agency (HTX), Singapore 138507. (email: daniel_teo@htx.gov.sg).

Z. Lin is with School of Electrical & Electronic Engineering, Nanyang Technological University (NTU), Singapore 639798, (e-mail: ezplin@ntu.edu.sg).

H. Zhu is with Singapore Institute of Manufacturing Technology (SIMTech), Agency for Science,Technology and Research (A*STAR), 2 Fusionopolis Way, Singapore, 138634, Singapore. (e-mail: zhu_haiyue@simtech.a-star.edu.sg).

† Corresponding author: zhu_haiyue@simtech.a-star.edu.sg

Language–Image Pre-training (CLIP) Image Encoder [4] with the object decoder branch. Moreover, our framework employs the decoded object feature as the prompt to orient the inference of both the grasp pose and relationship. Therefore, the prompt-based learning acts as an intrinsic bridge between the grasp pose and relationship branches to form a seamless integration. Moreover, the relationship detection is formulated as a relation matrix regression task. Compared with the traditional ROI-based solutions [5]–[7], our approach can avoid the traversal of all object pairs and outputs the entire relation in a single forward, which largely minimizes the computational burden. Overall, our framework is able to provide an optimal grasping execution with the awareness of both the manipulation order and grasp pose, thus enabling us to autonomously handle the objects in a reasonable and safe manner.

For most deep-learning-based robot solutions, the practical challenge also exists in the model generalization and adaptation capability for changing deployment environments. Unlike the standard computer vision tasks, those publicly available datasets for robotics are generally not universal enough to cover diverse real-world robot deployment scenarios. As a result, in this work, we propose a test-time adaptation scheme for *RelationGrasp*, which is based on the consistency-guided regularization. Therefore, such self-supervision technique facilitates the smooth adaptation to new deployment scenarios. The contributions of this paper are summarized as follows:

- We propose *RelationGrasp*, a unified framework for simultaneous open-vocabulary object detection, grasp detection, and manipulation relationship inference.
- A prompt-guided decoding mechanism to associate grasps with corresponding objects, bridging the grasp detection and relationship inference.
- A consistency-based self-supervised adaptation strategy for facilitating the model deployment in new environments.
- Our approach achieves the state-of-the-art in both manipulation relationship and grasp detection with a significant margin.

II. Related Works

A. Grasp Pose Detection

The realm of robotic grasp detection has been significantly impacted by data-driven methodologies in recent years. Both the 2D rectangle grasping [8]–[11] and 6-DoF pose grasping [12]–[14] have been extensively explored. To address the multi-object scenarios, both 2D heatmaps [10], [15], [16] and 3D equivalents [12], [17] are proposed to localize the graspness as the first stage, and then for corresponding pose regression in the second stage. The grasp detection is also treated as a special formation of object detection problem [3], [5], [18], and most works build the grasp detectors based on two-stages object detection models such as Faster-RCNN. Various sensor formats are explored as the input for the prediction network, e.g., RGB [10], [11], depth [10], [19], [20], RGB-D [10], [16], point cloud [12], [17], etc. Other methods take depth image as input for grasp inference, but they rely on additional sensors and their accuracy is highly dependent on the quality of the input depth image. Different from the previous approaches, our *RelationGrasp* detects the grasps based on a transformer encoder-decoder structure adapted from DETR, which does not rely on the anchor or heatmap for prediction.

B. Visual Manipulation Relationship

The manipulation relationship inference for objects in complex scenes is an essential task in robotics. Early approaches to object hierarchy inference rely on model-based techniques. For instance, point cloud [21] is utilized to infer object contact and interpret object hierarchies, and [22] employs a sparse Bayesian approach to learn spatial relationships using simulated objects. Data-driven approaches, particularly deep learning, have gained prominence in recent years. With those real-world or simulated datasets like VMRD [5], [23] and REGRAD [24], which contains object stacking scenes with relationship annotations, various data-driven methods for stacking hierarchy inference have been proposed. Convolutional Neural Networks (CNN)-based approaches, e.g., multi-task CNN [7] and VMRN [25], infers the inter-object relationship by treating them as a classification task for every object pair. GVMRN [26] employs a hybrid CNN-GNN network to infer the relation graph of all objects in a scene by integrating a graph neural network into the object detector. DUQIM-Net [27] builds upon the DETR object detector to directly predict the adjacency matrix of the relation graph.

III. Methodology

A. Overview

The objective of this work is to address the universal grasping problem for complex and challenging scenarios with highly stacked objects. Considering the object overlapping and clustering phenomenons, in order to complete the grasping task with minimal collision and damage, an intelligent robot solution needs to be able to reason the optimal grasping sequence as well as infer the suitable grasp pose simultaneously. Therefore, the *RelationGrasp* task in this work is formulated with three branches, i.e., 1) to detect and segment all instances in a scene with open-vocabulary capability, 2) to reason the relationship for all detected objects, and 3) to infer all suitable grasp candidates for objects. In this work, the proposed *RelationGrasp* is a novel end-to-end framework for 2D grasp prediction, which is anchor-less and simultaneously treats the object detection, grasp and relation prediction as a single joint problem. Unlike previous methods which rely on anchor-based or heatmap-based grasp box prediction, our model predicts all possible grasp boxes in a single forward pass, whose post-processing computation cost is significantly reduced compared to anchor-based or heatmap-based models.

Specifically, our framework, illustrated as in Fig. 1, consists of three branches, i.e., Object Branch, Relation Branch,

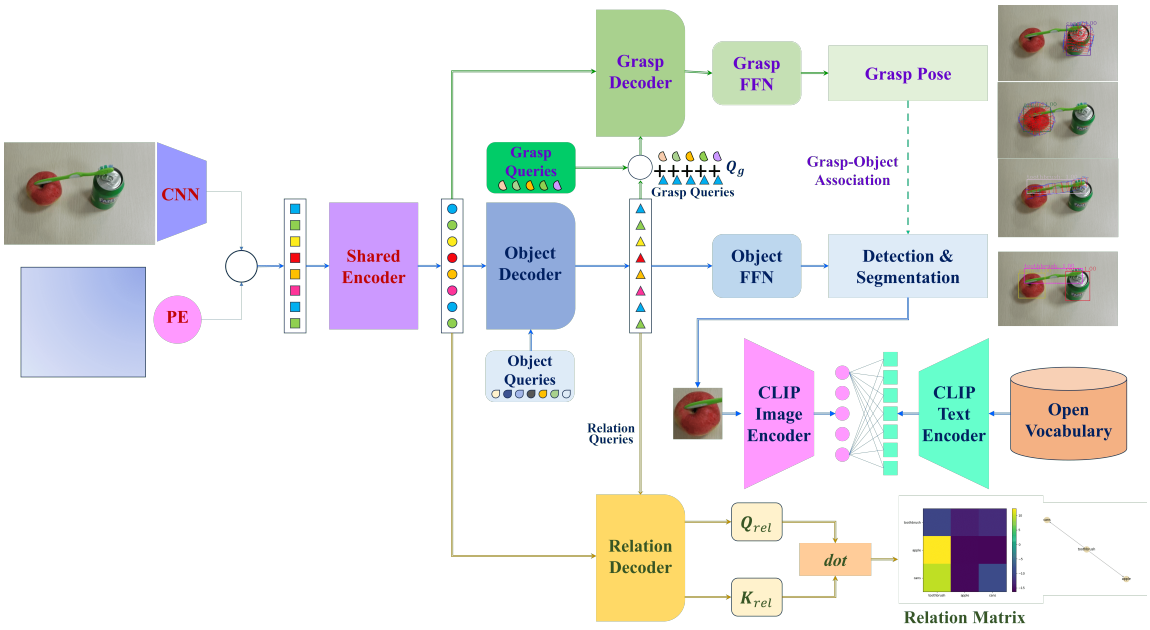

Fig. 1. *RelationGrasp*: The Shared Encoder takes the input feature map together with the Positional Embedding (PE), which is shared by all three parallel branches for simultaneous object detection, grasp pose detection, and manipulation relationship inference. The open-vocabulary object detection is achieved by integrating the CLIP Image/Text Encoders with the detection branch. Our framework employs the decoded object feature as the prompt to orient the prediction of grasp pose and manipulation relationship, which naturally bridges the association between the objects and grasp poses. Moreover, the relationship inference is achieved as a relation matrix regression task, which is handled by a relation-specific decoder. The proposed object-orientated multi-task learning framework greatly facilitate the adaption for novel deployment environments.

and Grasp Branch. The task of Object Branch is to detect all objects $\boldsymbol{O} = \{\boldsymbol{o}_i\}_{i=1}^{N}$, the Relation Branch is to reason the relationship $\boldsymbol{M}_{rel}(\boldsymbol{o}_i, \boldsymbol{o}_j)$ for all detected objects, and the Grasp Branch is to detect all possible grasp candidates $\boldsymbol{G} = \{\boldsymbol{g}_i\}_{i=1}^{M}$. In particular, we adopt the relation matrix $\boldsymbol{M}_{rel}$ to represent the relationship between objects, where $\boldsymbol{M}_{rel}(\boldsymbol{o}_i, \boldsymbol{o}_j)$ is a binary classification task, i.e., $\boldsymbol{o}_i$ is on $\boldsymbol{o}_j$ if $\boldsymbol{M}_{rel}(\boldsymbol{o}_i, \boldsymbol{o}_j)$ is larger than a threshold, or no relationship. Note that the relationship $\boldsymbol{o}_i$ is under $\boldsymbol{o}_j$ will be predicted from the symmetric element $\boldsymbol{M}_{rel}(\boldsymbol{o}_j, \boldsymbol{o}_i)$. In this work, we use the rectangle representation $\boldsymbol{g}_i = \big(x_i, y_i, w_i, h_i, \sin(\theta_i)\big)$ for grasp prediction, where (x_i, y_i) are the center coordinate of the grasp box, w_i is the required width of gripper plate, h_i is the gripper opening distance, and θ_i is the grasp pose orientation.

B. Network Architecture

Instead of using three separate branches for individual object detection, relationship, and grasp prediction tasks, a coupled-parallel-branch structure is designed in this work to promote learning efficiency through multi-task regularization.

1) Shared Backbone & Encoder: Taking an RGB image $\boldsymbol{I} \in \boldsymbol{R}^{H \times W \times 3}$, the feature for transformer encoder input is extracted by,

$$\boldsymbol{X}_{ei} = \mathbf{Flatten}\big(\mathcal{F}_b(\boldsymbol{I})\big), \tag{1}$$

where $\boldsymbol{X}_{ei} \in \mathbf{R}^{c \times H_0 W_0}$, $\mathcal{F}_b(\cdot)$ is the backbone feature extractor, and $\mathbf{Flatten}(\cdot)$ is the flatten operation, H and W are the height and width of the image. The sinusoidal positional encoding, PE, is then added to the feature embedding. The transformer encoder adopts the standard architecture comprising multiple layers, each layer containing a Multi-Head Self-Attention (MHSA) module and a Feed Forward Network (FFN) layer, denoted as,

$$\boldsymbol{X}_{eo} = \mathrm{TE}\big([\boldsymbol{x}_{ei}, PE]\big), \tag{2}$$

where $\boldsymbol{X}_{eo} \in \boldsymbol{R}^{d \times HW}$ will be the common feature for all decoder branches.

2) Open-Vocabulary Object Detection Branch: To improve the generalization capability of our model to unseen objects, we propose the open-vocabulary object detection in RelationGrasp, which allows to recognize open object categories described by the natural language. In this setting, the object detection module operates under a class-agnostic framework, where the primary goal is to identify regions of interest (ROIs) without assigning them to specific object categories during the initial detection decoding phase. The decoder for object detection branch follows the similar transformer design in DETR, where a learnable object query $\boldsymbol{Q}_o \in \mathbf{R}^{d \times N_q}$ is employed to serve as references to attend to specific instances. The class-agnostic object decoder, $\mathrm{TD}_o(\cdot)$, outputs the object feature as,

$$\boldsymbol{X}_{oo} = \mathrm{TD}_o(\boldsymbol{X}_{eo}, \boldsymbol{Q}_o) \in \mathbf{R}^{d \times N_q}. \tag{3}$$

The prediction of the bounding box use the corresponding FFNs.

Upon detection, the bounding boxes encapsulating the potential objects are cropped subsequently and then encoded using the CLIP [4], thus transforms into high-dimensional

feature vectors for open-vocabulary object classification. Simultaneously, a comprehensive vocabulary that includes a large list of potential object categories (open vocabulary) is encoded via the CLIP Text Encoder, which converts textual descriptions into their corresponding text features. The core of the open-vocabulary classification is the attention between the object image feature and the category text feature, which is simply achieved by dot product and softmax operation. As the two feature spaces are semantically aligned, it effectively assigns the most probable object category to each object region using the highest score.

This zero-shot classification approach allows the model to perform open-vocabulary object detection without the need for extensive labeled datasets. Instead, the model leverages the descriptive power of the CLIP encoder to generalize beyond the constraints of the available data. This is particularly advantageous for our robotic grasp detection task, where the necessity for parallel jaw grasp annotation limits the quantity of data that can be utilized for training. Through this methodology, RelationGrasp can accurately detect and classify a wide variety of objects using a relatively small training dataset, thereby demonstrating the feasibility of open-vocabulary object detection in resource-constrained applications.

3) Relation Branch: We formulate the relation detection problem as a prediction problem of the relation matrix. The relation decoder, $\mathrm{TD}_r(\cdot)$, takes the object feature from the object decoder as the relation query,

$$\boldsymbol{X}_{ro} = \mathrm{TD}_r(\boldsymbol{X}_{eo}, \boldsymbol{X}_{oo}) \in \mathbf{R}^{d\times N_q}. \tag{4}$$

Next, a simple self-attention mechanism is utilized to directly estimate the relation matrix,

$$\begin{aligned} \boldsymbol{Q}_{rel} &= \boldsymbol{W}_q \boldsymbol{X}_{ro},\ \boldsymbol{Q}_{rel} \in \mathbf{R}^{d\times N_q}, \\ \boldsymbol{K}_{rel} &= \boldsymbol{W}_k \boldsymbol{X}_{ro},\ \boldsymbol{K}_{rel} \in \mathbf{R}^{d\times N_q}, \\ \boldsymbol{M}_{rel} &= \sigma(\boldsymbol{Q}_{rel}^T \boldsymbol{K}_{rel}),\ \boldsymbol{M}_{rel} \in \mathbf{R}^{N_q\times N_q}, \end{aligned} \tag{5}$$

where $\boldsymbol{W}_q, \boldsymbol{W}_k \in \mathbf{R}^{d\times d}$ are the projection matrices, $\sigma(\cdot)$ is the sigmoid function, and $\boldsymbol{M}_{rel}$ is the relation matrix.

4) Grasp Branch: To associate the grasp candidates with specific detected objects, the grasp decoder is designed with object-aware query $\boldsymbol{Q}_g$, which encodes both the object association information in the image level and the positional information in the object level. In the image level, the object-wise decoded feature vector $\boldsymbol{x}_{oo}^i \in \mathbf{R}^d$ from $\boldsymbol{X}_{oo} \in \mathbf{R}^{d\times N_q}$ is utilized to encode the object association for query. Assume N_g grasp boxes are predefined to be predicted in each object. In the object level, an additional randomly-initialized learnable query embedding $\boldsymbol{Q}_{gp}^i \in \mathbf{R}^{d\times N_g}$ is employed to encode the positional difference among the grasps and,

$$\boldsymbol{Q}_g^i = \boldsymbol{Q}_{gp}^i + [\boldsymbol{x}_{oo}^i]^{N_g} \in \mathbf{R}^{d\times N_g} \tag{6}$$

where $[\,\cdot\,]^{N_g}$ denotes the broadcast operation with dimension N_g.

In the grasp branch, the decoder $\mathrm{TD}_g(\cdot)$ takes the grasp query $\boldsymbol{Q}_g$ for each object,

$$\boldsymbol{X}_{go}^i = \mathrm{TD}_g(\boldsymbol{X}_{eo}, \boldsymbol{Q}_g^i) \in \mathbf{R}^{d\times N_q}. \tag{7}$$

The final grasp estimation can be treated as a regression task,

$$\begin{aligned} \boldsymbol{G}^i &= \sigma(\boldsymbol{W}_g \boldsymbol{X}_{ro}^i), \\ \boldsymbol{C}_g^i &= \mathrm{softmax}(\boldsymbol{W}_{gc} \boldsymbol{X}_{ro}^i), \end{aligned} \tag{8}$$

where $\boldsymbol{G}^i = \{\boldsymbol{g}_j^i\}_{j=1}^{N_g}$ and $\boldsymbol{C}_g^i = \{\boldsymbol{c}_j^i\}_{j=1}^{N_g}$ denote the predicted grasps and its graspable indicator, respectively. $\boldsymbol{W}_g \in \mathbf{R}^{5\times d}$ and $\boldsymbol{W}_{gc} \in \mathbf{R}^{2\times d}$ are the respective projection matrices.

5) Fully-Supervised Losses: The losses are designed separately for each branch.

For the grasp branch, the ground-truth grasps are represented as

$$\widehat{\boldsymbol{G}}_{gt} = \big\{\{\hat{\boldsymbol{g}}_j^i, \hat{c}_j^i\}_{j=1}^{N_g}\big\}_{i=1}^{N_q}, \tag{9}$$

where $\hat{c}_j^i = 1$ or $\emptyset$ is introduced to suit the fixed lengths of both object-level N_q and grasp level N_g. The below Hungarian loss is employed to obtain the supervision loss for the grasp branch,

$$\begin{aligned} \mathcal{L}_g = \sum_{i=1}^{N_q}\sum_{j=1}^{N_g} \Big[&-\log p_{\sigma_i(j)}(\hat{c}_j^i) \\ &+ \mathbb{1}_{\{\hat{c}_j^i \neq \emptyset\}} \mathcal{L}_{rbox}\big(\hat{\boldsymbol{g}}_j^i, \boldsymbol{g}_{\sigma_i}(j)\big)\Big], \end{aligned} \tag{10}$$

where $\sigma_i(j)$ denotes the optimal matching from the predicted grasps to j-th ground truth in i-th object, $p_\sigma(\hat{c})$ denotes the predicted probability to be class $\hat{c}$. $\mathcal{L}_{rbox}(\cdot)$ represents the loss for grasp boxes, which includes both the element-wise l_1 loss and orientation-aware IoU loss.

For the relation branch, considering the sparsity in the ground-truth relation matrix $\widehat{\boldsymbol{M}}_{rel}$, the element-wise classification is with high-class imbalance. Therefore, the focal loss is adopted for the relation branch,

$$\mathcal{L}_{rel} = -\sum_{i,j} \widehat{\boldsymbol{M}}_{rel}(i,j)\big(1 - \boldsymbol{M}_{rel}(i,j)\big)^\gamma \log\big(\boldsymbol{M}_{rel}(i,j)\big), \tag{11}$$

where γ is a parameter. For the object branch, its loss $\mathcal{L}_{obj}$ simply follows the original DETR. Finally, the fully-supervised loss is

$$\mathcal{L}_{fs} = \lambda_{obj}\mathcal{L}_{obj} + \lambda_{rel}\mathcal{L}_{rel} + \lambda_g \mathcal{L}_g, \tag{12}$$

where λ_{obj}, λ_{rel}, and λ_g are the respective weights.

C. Consistency-Based Self-Supervised Adaptation

In this section, we present a consistency-based self-supervised adaptation to enhance the inference performance in a novel environment or domain. The objective is to adapt the *RelationGrasp* model from the source domain ($\mathcal{D}_s$ with plenty of annotations) to a new target domain dataset $\mathcal{D}_u$ that only contains partial annotations, e.g., object bounding box. We employ a Teacher-Student learning structure to facilitate the adaption. The teacher model, denoted as $\mathcal{F}_T(\cdot)$, is fully supervised from the source dataset $\mathcal{D}_s$ as well as the annotated portion of $\mathcal{D}_u$. The consistency is emphasized on the student model $\mathcal{F}_S(\cdot)$ using the pseudo labeling generated by $\mathcal{F}_T(\cdot)$.

Specifically, given a sample $\boldsymbol{I}_u \in \mathcal{D}_u$ with partial or no label, a weak augmentation $Aug_w(\cdot)$ is applied to obtain the pseudo label $\mathcal{F}_T\big(Aug_w(\boldsymbol{I}_u)\big)$, and the strong augmentation $Aug_s(\cdot)$ is applied on the inference sample for $\mathcal{F}_S(\cdot)$. The consistency regularization is to ensure the prediction agreements between the strong and weak augmentations, i.e.,

$$\mathcal{L}_{ss} = \mathcal{L}\Big(\mathcal{F}_S\big(Aug_s(\boldsymbol{I}_u)\big), {}^{s}T_w\big(\mathcal{F}_T\big(Aug_w(\boldsymbol{I}_u)\big)\big)\Big), \quad (13)$$

where ${}^{s}T_w(\cdot)$ is the geometry transform from weak augmentation to strong augmentation. Notably, since the new domain features unseen objects, we ensure adaptability by adopting a class-agnostic object detection branch. The teacher model is updated through an Exponential Moving Average (EMA) from the online student model with a coefficient of 0.999.

IV. Experiment

To validate the performance of the proposed algorithm, we evaluate our model on three publicly available datasets, i.e., VMRD [23], OCID-grasp [28] and MetaGraspNet [29].

A. Implementation Details

Our network is implemented using the PyTorch framework. Our transformer hidden dimension d is set to 256, the number of grasp decoder queries N_g is set to 20 for efficient decoding, and the number of object decoder queries N_q is set to 100. The image encoder, object decoder, and grasp decoder each consist of 6 layers, while the relation decoder has 2 layers. All transformer modules use 8 attention heads. To mitigate overfitting, we apply random augmentations including flips, resizing, color jittering, Gaussian blur, and rotation to both images and annotations, for all three datasets. All models are trained using two RTX 3090 GPUs. We initiate training from a pretrained DETR-DC50 model with a ResNet-50 backbone trained on the COCO2017 dataset. We employ the AdamW optimizer with learning rates of 1×10^{-4} for transformer modules and 1×10^{-5} for the ResNet backbone. The learning rate decays by a factor of 10 every 300 epochs.

B. Evaluation Metrics

For the grasp evaluation, in line with prior work, we employ the (o, g) metric proposed by [5]. For evaluating the relationship inference, we adopt three standard benchmarks [23], i.e., Object Precision (OP), Object Recall (OR), and Image Accuracy (IA).

C. Results on VMRD Dataset

The Visual Manipulation Relationship Detection (VMRD) [5], [23] dataset consists of 4233 training samples and 450 test samples, which encompasses 31 object categories, and each image contains 2 to 5 stacked objects.

Grasp Pose Accuracy: Table I shows the results of our grasp detection performance on the VMRD dataset. Our *RelationGrasp* surpasses the performance of previous Faster-RCNN-based grasp detectors in multi-object scenarios. This superiority can be attributed to the intrinsic advantages of object prompt-based grasp pose detection, which excels at accurately associating the grasps with correct objects. In contrast, those traditional IoU-based grasp-object matching techniques [5], [30] often struggle when dealing with tightly clustered objects, which leads to grasp association errors in densely stacked scenarios. Consequently, our model, adopting the feature representations as the queries of detected objects, genuinely understands the object context in decoding grasp poses, leading to higher multi-object grasping success rates.

TABLE I
GRASP POSE ACCURACY ON VMRD DATASET (%)

Methods	mAPg
Faster RCNN + FCGN	54.5
ROI-GD [5]	68.2
Zhang [7]	70.5
Keypoint-based scheme [6]	74.3
Multi-Stage ROI Extraction [31]	74.8
MMD [32]	76.7
RelationGrasp (ours)	**79.3**

TABLE II
RELATIONSHIP ACCURACY ON VMRD DATASET (%)

Methods	OR	OP	IA
Multitask CNN [7]	86.0	88.8	67.1
VMRN-RN101 [25]	85.4	85.5	65.8
VMRN-VGG16 [25]	86.3	88.8	68.4
GVMRN-RF-RN101 [26]	86.9	87.5	68.8
GVMRN-RF-VGG16 [26]	88.7	89.5	70.2
GGNN-VMRN-RN101 [30]	90.09	88.01	75.33
GGNN-VMRN-VGG16 [30]	89.64	88.00	75.56
Adj-Net RN50 [27]	88.9	91.5	75.0
Adj-Net RN101 [27]	89.8	91.5	77.3
RelationGrasp (ours)	**97.1**	**95.7**	**83.8**

Relationship Accuracy: Table II shows the results in relationship inference against the baselines on the VMRD dataset. For a more nuanced assessment, Table III offers a detailed breakdown of relationship inference accuracy under scenes with a specific number of objects. In both cases, *RelationGrasp* achieves the new state-of-the-art accuracy in relationship inference. Prior methods [23], [25], [30] relying on Faster-RCNN architecture, determine the relationships by pooling features from individual proposals and the union area between object pairs. Unfortunately, these approaches do not effectively consider the global information and scene context, causing performance degradation in complex scenarios.

D. Results on OCID-grasp Dataset

The OCID-grasp dataset [28] is an extension of the OCID dataset [33] with additional grasp pose and category annotations for graspable objects, which includes 1,763 images, 11.4k objects, and over 75k hand-annotated grasp candidates with objects categorized into 31 classes.

Grasp Pose Accuracy: Table IV shows the results of the grasp pose detection performance on the OCID-grasp dataset. Following previous works [28], [34], we adopt grasp accuracy as the evaluation metric, which shows that our results achieve state-of-the-art performances.

TABLE III

RELATIONSHIP ACCURACY WITH SPECIFIED OBJECT NUMBERS ON VMRD DATASET (%)

Methods	IA-2	IA-3	IA-4	IA-5
Multitask CNN [7]	87.7	64.1	56.6	72.9
VMRN-RN101 [25]	80.00	58.37	47.17	54.29
VMRN-VGG16 [25]	86.15	61.72	62.26	64.29
GVMRN-RF-RN101 [26]	91.4	69.2	61.2	57.5
GVMRN-RF-VGG16 [26]	92.9	70.3	63.8	60.3
GGNN-VMRN-RN101 [30]	92.31	75.12	66.98	72.86
GGNN-VMRN-VGG16 [30]	86.15	73.21	69.81	82.86
Adj-Net RN50 [27]	87.3	74.5	69.8	72.6
Adj-Net RN101 [27]	88.7	75.2	75.0	76.7
RelationGrasp (ours)	**95.4**	**79.9**	**79.2**	**91.4**

TABLE IV

GRASP POSE ACCURACY ON OCID-GRASP DATASET (%)

Methods	Grasp Accuracy
GGCNN-2 [35]	63.4
GR-ConvNet [15]	74.1
EfficientGrasp RGB-D [36]	76.4
Det-Seg-Refine [28]	89.02
GSMR-CNN [34]	91.9
RelationGrasp (Ours)	**92.21**

Domain Adaptation Performance: We have also evaluated the effectiveness of the proposed consistency-based domain adaptation method using the OCID-grasp dataset. Firstly, a class-agnostic *RelationGrasp* model is trained on VMRD dataset using both object bounding box and grasp pose annotations, we evaluate this pre-trained model on OCID-grasp dataset as the baseline without adaptation, denoted as VMRD-Pretrained Model in Table V. Secondly, using the proposed self-supervised adaptation, we adapt the pre-trained model, where no grasp pose annotations are utilized from the OCID-grasp dataset. We denote this as Semi-Supervised Adaptation in Table V. Lastly, to provide the performance upper bound information, we also provide the results of a fully-supervised model on OCID-grasp dataset using all available annotations, denoted as Fully-Supervised Upper Bound. The experiment results show that even without using any grasp pose annotations in new environments, our self-supervised adaptation approach can achieve significant improvements in grasp accuracy at almost free cost.

E. Results on MetaGraspNet Dataset

We also conduct experiments on the MetaGraspNet dataset [29], which is a recent dataset featuring multi-task learning including grasp pose detection and relationship inference. We employ a class-agnostic model pre-trained on the VMRD dataset as the reference, denoted as VMRD-Pretrained Model. Subsequently, we fine-tune this model on the MetaGraspNet dataset, denoted as Fine-Tuned on MetaGraspNet, and both results are presented in Table VI. It demonstrates that, despite the small training set size, *RelationGrasp* delivers promising fine-tuning adaptation performances.

TABLE V

DOMAIN ADAPTATION RESULTS ON OCID-GRASP DATASET (%)

Methods	Grasp Accuracy	mAPg
VMRD-Pretrained Model	39.28	16.26
Semi-Supervised Adaptation	55.07	29.81
Fully-Supervised Upper Bound	92.21	76.1

TABLE VI

EVALUATION RESULTS ON METAGRASPNET (%)

Metrics	VMRD-Pretrained Model	Fine-Tuned on MetaGraspNet
mAP	24.63	79.62
mAPg	9.14	50.08
OR	19.88	70.30
OP	74.05	81.23
IA	18.68	50.08
IA-2	28.1	88.46
IA-3	28.91	80.0
IA-4	0.0	26.32

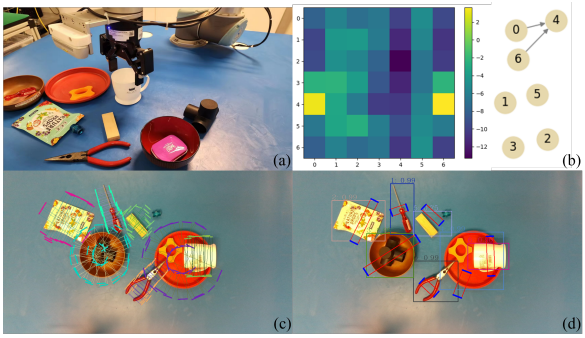

Fig. 2. Experiment setup and results of novel scene and objects: (a) robot setup, (b) detected relationship, (c) detected grasp poses, and (d) detected objects with best grasps.

F. Results of Real-World Experiments

Finally, the real robot experiment is conducted to verify the practical usage for deployment with novel objects and scenes. Fig. 2(a) shows the experiment setup using a Universal Robot UR5, an RGB-D camera Realsense D435 mounting on hand to capture the online scene images, and the end-effector being a Robotiq 2F-85 gripper. We select 20 unseen (not in training dataset) objects to evaluate the real grasping performance, where Fig. 2(b) shows all the detected grasp poses from one scene, and Fig. 2(c) indicates the detected objects with their respective best grasps (highest score), and Fig. 2(d) plots the detected relationship among the objects. In the experiment, we use the detected relationship information to guide the object grasping sequence, and for each object, we take its best grasp for robot execution. Our experiments show that the proposed *Relationship* can achieve an average success rate of 72.7% for real grasping, which indicates the practical usages of our framework in real-world applications. More demo information can be found in the supplementary video.

V. CONCLUSION

This paper presents *RelationGrasp*, an innovative framework for integrated grasp sequence planning and grasp pose

detection as well as open-vocabulary object detection. Our experimental results illustrate the effectiveness of prompt-guided grasp detection, which accurately associates predicted grasps with their corresponding graspable objects, notably enhancing the grasp accuracy, particularly in complex multi-object scenarios. Moreover, by framing inter-object relationships as an relation matrix regression task, *RelationGrasp* also establishes a new state-of-the-art benchmark in manipulation relationship inference.

REFERENCES

[1] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 1316–1322.

[2] S. Kumra and C. Kanan, "Robotic grasp detection using deep convolutional neural networks," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 769–776.

[3] F.-J. Chu, R. Xu, and P. A. Vela, "Real-world multiobject, multigrasp detection," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3355–3362, 2018.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[5] 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.

[6] T. Li, F. Wang, C. Ru, Y. Jiang, and J. Li, "Keypoint-based robotic grasp detection scheme in multi-object scenes," *Sensors*, vol. 21, no. 6, p. 2132, 2021.

[7] 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 *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 6435–6442.

[8] D. Wang, C. Liu, F. Chang, N. Li, and G. Li, "High-performance pixel-level grasp detection based on adaptive grasping and grasp-aware network," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 11, pp. 11 611–11 621, 2022.

[9] H. Cheng, Y. Wang, and M. Q.-H. Meng, "A robot grasping system with single-stage anchor-free deep grasp detector," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[10] S. Wang, Z. Zhou, and Z. Kan, "When transformer meets robotic grasping: Exploits context for efficient grasp detection," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 8170–8177, 2022.

[11] L. Chen, P. Huang, Y. Li, and Z. Meng, "Edge-dependent efficient grasp rectangle search in robotic grasp detection," *IEEE/ASME Transactions on Mechatronics*, vol. 26, no. 6, pp. 2922–2931, 2021.

[12] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 441–11 450.

[13] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Transactions on Robotics*, pp. 1–17, 2023.

[14] Z. Liu, Z. Chen, S. Xie, and W. Zheng, "Transgrasp: A multi-scale hierarchical point transformer for 7-dof grasp detection," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 1533–1539.

[15] S. Kumra, S. Joshi, and F. Sahin, "Antipodal robotic grasping using generative residual convolutional neural network," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 9626–9633.

[16] H. Zhu, Y. Li, F. Bai, W. Chen, X. Li, J. Ma, C. S. Teo, P. Yuen Tao, and W. Lin, "Grasping detection network with uncertainty estimation for confidence-driven semi-supervised domain adaptation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 9608–9613.

[17] C. Wang, H.-S. Fang, M. Gou, H. Fang, J. Gao, and C. Lu, "Graspness discovery in clutters for fast and accurate grasp detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 15 944–15 953.

[18] X. Zhou, X. Lan, H. Zhang, Z. Tian, Y. Zhang, and N. Zheng, "Fully convolutional grasp detection network with oriented anchor box," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 7223–7230.

[19] G. Chalvatzaki, N. Gkanatsios, P. Maragos, and J. Peters, "Orientation attentive robotic grasp synthesis with augmented grasp map representation," *arXiv preprint arXiv:2006.05123*, 2020.

[20] D. Morrison, P. Corke, and J. Leitner, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," *arXiv preprint arXiv:1804.05172*, 2018.

[21] B. Rosman and S. Ramamoorthy, "Learning spatial relationships between objects," *The International Journal of Robotics Research*, vol. 30, no. 11, pp. 1328–1342, 2011.

[22] K. Sjöö and P. Jensfelt, "Learning spatial relations from functional simulation," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2011, pp. 1513–1519.

[23] 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.

[24] 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.

[25] H. Zhang, X. Lan, X. Zhou, Z. Tian, Y. Zhang, and N. Zheng, "Visual manipulation relationship recognition in object-stacking scenes," *Pattern Recognition Letters*, vol. 140, pp. 34–42, 2020.

[26] G. Zuo, J. Tong, H. Liu, W. Chen, and J. Li, "Graph-based visual manipulation relationship reasoning in object-stacking scenes," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[27] 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.

[28] S. Ainetter and F. Fraundorfer, "End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from rgb," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 452–13 458.

[29] M. Gilles, Y. Chen, T. R. Winter, E. Z. Zeng, and A. Wong, "MetaGraspnet: A large-scale benchmark dataset for scene-aware ambidextrous bin picking via physics-based metaverse synthesis," in *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2022, pp. 220–227.

[30] 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.

[31] J. Xia, J. Chi, C. Wu, and F. Zhao, "Robot grasping detection in object overlapping scenes based on multi-stage roi extraction," in *2022 34th Chinese Control and Decision Conference (CCDC)*. IEEE, 2022, pp. 5066–5071.

[32] X. Dong, Y. Jiang, F. Zhao, and J. Xia, "A practical multi-stage grasp detection method for kinova robot in stacked environments," *Micromachines*, vol. 14, no. 1, p. 117, 2022.

[33] M. Suchi, T. Patten, D. Fischinger, and M. Vincze, "Easylabel: A semi-automatic pixel-wise object annotation tool for creating robotic rgb-d datasets," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 6678–6684.

[34] V. Holomjova, A. J. Starkey, and P. Meißner, "GSMR-CNN: An end-to-end trainable architecture for grasping target objects from multi-object scenes," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 3808–3814.

[35] D. Morrison, P. Corke, and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *The International journal of robotics research*, vol. 39, no. 2-3, pp. 183–201, 2020.

[36] H. Cao, G. Chen, Z. Li, Q. Feng, J. Lin, and A. Knoll, "Efficient grasp detection network with gaussian-based grasp representation for robotic manipulation," *IEEE/ASME Transactions on Mechatronics*, 2022.