2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

SRI-Graph: A Novel Scene-Robot Interaction Graph for Robust Scene Understanding

Dong Yang, Xiao Xu, Mengchen Xiong, Edwin Babaians, and Eckehard Steinbach

***Abstract*— We propose a novel scene-robot interaction graph (SRI-Graph) that exploits the known position of a mobile manipulator for robust and accurate scene understanding. Compared to the state-of-the-art scene graph approaches, the proposed SRI-Graph captures not only the relationships between the objects, but also the relationships between the robot manipulator and objects with which it interacts. To improve the detection accuracy of spatial relationships, we leverage the 3D position of the mobile manipulator in addition to RGB images. The manipulator's ego information is crucial for a successful scene understanding when the relationships are visually uncertain. The proposed model is validated for a real-world 3D robot-assisted feeding task. We release a new dataset named 3DRF-Pos for training and validation. We also develop a tool, named LabelImg-Rel, as an extension of the open-sourced image annotation tool LabelImg for a convenient annotation in robot-environment interaction scenarios*. Our experimental results using the Kinova® Movo platform show that SRI-Graph outperforms the state-of-the-art approach and improves detection accuracy by up to 9.83%.**

I. Introduction

The concept of visual scene understanding has recently gained considerable attention in robotics research, which is inspired by humans' ability to visually interpret and comprehend scenes effortlessly [1], [2]. It focuses not only on recognizing and localizing the objects present in a scene, but also understanding the relationships (*e.g.* spatial relationships) among them. This comprehensive understanding is of high interest for many applications in autonomous driving [3]–[5] and in robotics, such as robotic surgery, navigation and decision making [6]–[15]. The objective of scene understanding is to extract valuable information from the scene, much like humans do. However, current scene understanding research is still facing challenges for complex human-robot interaction (HRI) scenarios due to its low identification rate [16].

A fundamental task in scene understanding is scene graph generation (SGG), which comprises detecting object instances and their relationships in an image using the

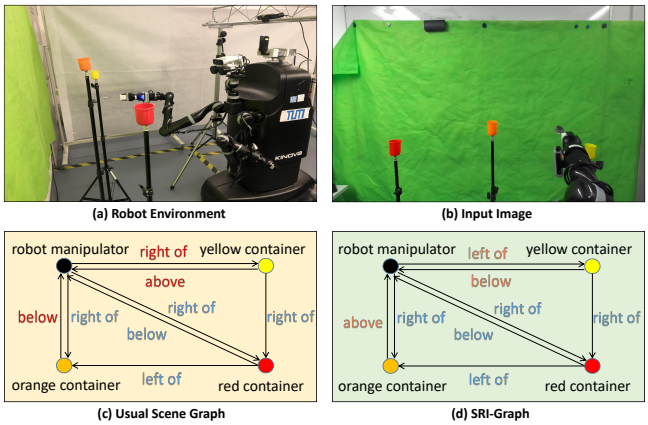

Fig. 1: An SRI-Graph example in the robot-assisted feeding task scenario using Kinova® Movo platform. (a) The robot environment. (b) An RGB image captured with an external camera. (c) Scene graph generated from the RGB image shown in (b) using the conventional model [17]. The incorrect detected relationships are marked in red. (d) Our SRI-Graph shows a robust and reliable performance, especially in case the position of the manipulator cannot be clearly determined from the image.

$\langle subject, predicate, object \rangle$ format. With a single RGB image as input, SGG models [16]–[23] generate a graph representation for the scene, where the nodes correspond to object bounding boxes with their labels, and the edges indicate their pairwise relationships [1] (see Figure 1). However, due to the limitation of existing datasets (*e.g.* Visual Genome (VG) [24] and GQA dataset [25]), most studies in this area consider only daily life scenes where robots are not present.

The authors in [26], [27] use scene understanding information from generated scene graphs to provide contextual information to a robot planner. Although they investigate the effect of scene understanding on the robotic task, they ignore the ego information of the robot in scenes. Their SGG model relies on pure RGB image-based information, for example, object labels, visual feature maps, and 2D bounding boxes. When a robot interacts with a scene, ego information (*e.g.* the 3D position of the robot manipulator) helps to distinguish the positional relationships between the manipulator and objects.

In this paper, considering a robot is involved in the scene, we propose a novel scene-robot interaction graph (SRI-Graph) based on SGG [16] that takes advantage of the known 3D position of the robot manipulator to enhance scene understanding. As the 3D positions can be easily and accurately read for most robot manipulators, our SRI-Graph can be applied widely in many robot-related task scenes. Figure 1 illustrates the setup used in this

All Authors are with the School of Computation, Information and Technology, Department of Computer Engineering, Chair of Media Technology (LMT) and Munich Institute of Robotics and Machine Intelligence (MIRMI), Technical University of Munich (TUM), Germany. - {dong.yang, xiao.xu, mengchen.xiong, edwin.babaians, eckehard.steinbach}@tum.de

The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany in the programme of "Souverän. Digital. Vernetzt.". Joint project 6G-life, project identification number: 16KISK002. Mengchen Xiong is supported by the Chinese Scholarship Council (CSC), Grant #202006290010.

* The dataset, annotation tool, and our framework will be public available upon acceptance.

paper. We consider a real-world robot-assisted feeding task in which the robot needs to pour food into the desired container. Our SRI-Graph is able to provide robust and reliable performance, especially in case the position of the manipulator cannot be accurately determined from the image. For example, when the manipulator is a little below the orange container, the scene graph based on [17] detects $\langle$*orange container*, *below*, *robot manipulator*$\rangle$, whereas the proposed SRI-Graph makes a correct detection (see Figure 1).

The training and validation are conducted based on our 3DRF-Pos dataset that is created according to the real-world feeding task. The 3DRF-Pos dataset consists of 760 RGB images where each image has an average of 3 objects, 1 robot manipulator, and 12 pairwise relationships among them and the corresponding 3D positions of the robot manipulator. In addition, we also developed a new relationship annotation tool LabelImg-Rel, as an extension of the publicly available image annotation tool LabelImg [28]. Both the 3DRF-Pos dataset and the annotation tool LabelImg-Rel will be open-sourced upon acceptance.

The main contributions of this paper are summarized as follows:

- Proposal of a novel and adaptable scene graph generation model by leveraging 3D positions of the robot manipulator as side information for robot-involved scenes to improve detection accuracy.
- Release of a new dataset for real-world 3D robot-assisted feeding tasks, 3DRF-Pos, and development of a new relationship annotation tool, LabelImg-Rel.
- Validation of the proposed SRI-Graph model in a real robot-assisted feeding task and performance comparison with the state-of-the-art benchmark.

The remainder of this paper is organized as follows. In Section II, we briefly review the related works. Section III discusses the problem statement. In Section IV, we describe the proposed SRI-Graph in detail. Section V introduces 3DRF-Pos, LabelImg-Rel, and the experiments. Section VI concludes this paper and discusses future work.

II. Related Works

A. Scene Understanding in Robotics

Previous studies have leveraged scene understanding information to improve the performance of robot-related tasks. Humblot-Renaux *et al*. in [7] presented a navigation-oriented framework to learn pixel-wise driveability from images for outdoor robotic navigation. Their model segmented the outdoor scenes into three different driveability labels (preferable, possible, and impossible) and helped the robot to decide how to navigate in the area with these labels. In [9], Wellhausen *et al*. proposed a weakly supervised and self-supervised learning approach to predict terrain characteristics from RGB images. This information was used for the path planning of a legged robot.

In addition to navigation, scene information was also applied in teleoperation tasks for guidance and object search.

TABLE I: Comparison of SRI-Graph with the state-of-the-art models

Criteria	Chandan *et al*. [12]	Tang *et al*. [16]	Amiri *et al*. [26]	SRI-Graph (Ours)
Object det.	✓	✓	✓	✓
Relation det.	✗	✓	✓	✓
Robot involved	✓	✗	✓	✓
Robot ego info	✗	✗	✗	✓

Chandan *et al*. in [12] developed a reinforcement learning-based framework (GHAL360) to help the teleoperated robot understand remote environments. It is able to analyze 360° visual scene information and actively guide human attention toward the object of interest. In [13], Zeng *et al*. proposed a visual object search strategy using the Semantic Linking Maps (SLiM) model. The robot was guided to explore promising areas that potentially contain the target object by updating the next best view pose iteratively. Kunze *et al*. in [14] developed a method to search for the target object based on the Qualitative Spatial Relations (QSRs) between it and landmark objects in the background. By utilizing the QSR scene descriptions, the search efficiency and accuracy were improved.

B. Scene Graph Generation (SGG)

Recently, Scene Graph Generation (SGG) has received increased interest. SGG extracts a comprehensive semantic representation of the scene and describes the relationships between objects [29]. The release of the VG dataset [24] has motivated a great variety of state-of-the-art approaches toward developing better feature extraction networks for enhancing the performance of SGG models. IMP [19], MSDN [20], and Graph R-CNN [17] integrate contextual information by applying recurrent neural networks (RNNs) and an attentional graph convolutional network (aGCN). Neural Motifs [21] analyzed the regularly appearing structures in scene graphs and employed bidirectional Long Short-Term Memories (LSTMs) [30] to determine the global context. Tang *et al*. in [16] investigated the use of causal Total Direct Effect (TDE) inference to achieve unbiased prediction from a biased annotated training dataset.

Although SGG has gained increasing attention, the number of studies that apply SGG on robotic tasks is still quite limited. Amiri *et al*. in [26] proposed visual scene analysis for robot planning (SARP) to provide the robot with contextual information. In [27], Li *et al*. explored the use of scene graphs in human-robot collaboration for a disassembly task scenario.

The aforementioned approaches neglect the ego information of the robot while acquiring the scene understanding information. Motivated by this, we propose a novel SRI-Graph to enhance the detection accuracy by considering the 3D position of the robot manipulator as side information. In

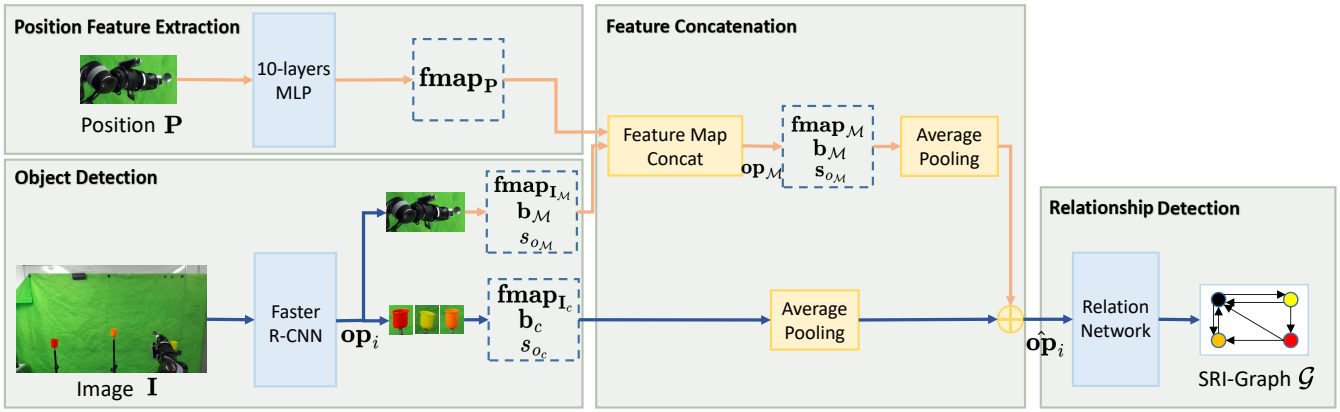

Fig. 2: SRI-Graph Generation Model Diagram. Our model consists of: Object detection; Position feature extraction; Feature concatenation and relationship detection network.

Table I, a comparison of SRI-Graph with other state-of-the-art models is summarized from diverse perspectives.

III. Problem Statement

Let $\boldsymbol{I}$ be the RGB image captured by the camera at a certain time. $\boldsymbol{P}$ denotes the corresponding 3D position of the end-effector of the robot manipulator in the robot base coordinate frame. Since both x, y, and z coordinates are relevant in most cases where the robot base is not parallel with the scene plane because of the physical constraints of the manipulator, we use 3D positions even though our RGB images are only two-dimensional. Given $\boldsymbol{I}$ and $\boldsymbol{P}$, we aim to generate a robust and reliable SRI-Graph $\mathcal{G}$ including rich scene understanding information, *i.e.* detected object labels (index $1, \ldots, m$), the robot manipulator (index $\mathcal{M}$), their bounding boxes, and n spatial relationships between them. The three components of the SRI-Graph $\mathcal{G}$ are noted as below:

- A set $\mathcal{O} = \{o_i | i = 1, \ldots, m, \mathcal{M}\}$ of object labels, and $o_i \in \mathcal{C}_o$, where $\mathcal{C}_o$ indicates the predefined object classes including m ordinary objects and the robot manipulator $o_{\mathcal{M}}$. Note that we consider only one robot manipulator in this paper.
- A corresponding set $\mathcal{B} = \{\boldsymbol{b}_i | i = 1, \ldots, m, \mathcal{M}\}$ of bounding boxes for $\mathcal{O}$. $\boldsymbol{b}_i = \{x_i, y_i, w_i, h_i\} \in \mathbb{R}^4$, where x_i, y_i, w_i, and h_i are the x-, y-coordinate, width, and height of the bounding box, respectively.
- A set $\mathcal{R} = \{\boldsymbol{r}_k | k = 1, \ldots, n\}$ of pairwise relationships among the elements of $\mathcal{O}$, where $r_k = \{(o_i, b_i), p_{i \to j}, (o_j, b_j)\}$ in the form of $\langle subject, predicate, object \rangle$ represents the relationship from the i-th to j-th object. $p_{i \to j} \in \mathcal{C}_{\mathcal{P}}$ is the predicate, $\mathcal{C}_{\mathcal{P}}$ denotes the predefined predicate classes. We set $\mathcal{C}_{\mathcal{P}} = \{$left of, right of, above, below$\}$ in our experiments.

IV. Scene Robot Interaction Graph

In this section, we first introduce the four stages of the SRI-Graph generation model: 1) Object Detection, 2) Position Feature Extraction, 3) Feature Concatenation, and 4) Relationship Detection. We then introduce our convenient relationship annotation tool, LabelImg-Rel. Afterward, we present our recorded dataset, including the RGB images and corresponding 3D positions of the robot manipulator (3DRF-Pos). An overview of our model is shown in Figure 2.

A. Object Detection Module

We apply Faster R-CNN [31], which is one of the most popular deep learning-based object detection systems, to detect objects and the robot manipulator in the scene. The input to Faster R-CNN is an RGB image $\boldsymbol{I}$ and the output is a set of predicted object proposals. Each object proposal $\boldsymbol{op}_i$ contains a visual feature map $\boldsymbol{fmap}_{\boldsymbol{I}_i}$, a bounding box $\boldsymbol{b}_i$, and a corresponding object label score $s_{o_i} \in (0, 1)$:

$$\boldsymbol{op}_i := \{\boldsymbol{fmap}_{\boldsymbol{I}_i}, \boldsymbol{b}_i, s_{o_i}\}. \tag{1}$$

where $i = 1, \ldots, m, \mathcal{M}$. $\boldsymbol{op}_{\mathcal{M}}$ represents for the robot and $\boldsymbol{op}_{\boldsymbol{I}_c}$ represent for other objects except the robot ($c = 1, \ldots, m$). We consider the label with the highest score among s_{o_i} as the detected object label o_i:

$$o_i := \underset{o_i \in \mathcal{C}_o}{\arg\max}\, s_{o_i}. \tag{2}$$

B. Position Feature Extraction From Robot

Previous work considers the robot in the scene as an ordinary object [7], [9], [12]–[14], [26], [27]. Only visual features extracted from the RGB images are applied and the robot ego information is ignored. Different from those studies, in our system we record the 3D positions of the robot manipulator in the robot base coordinate frame ($\boldsymbol{P} = [x, y, z]$) and provide a robot position feature map $\boldsymbol{fmap}_{\boldsymbol{P}}$ in addition to the visual feature map $\boldsymbol{fmap}_{\boldsymbol{I}_i}$.

To do that, we observe the timestamp for each position record in order to synchronize it with the captured RGB image $\boldsymbol{I}$. To extract $\boldsymbol{fmap}_{\boldsymbol{P}}$ of $\boldsymbol{P}$, we employ a multi-layer perceptron (MLP) consisting of 10 fully connected layers in the residual block structure. It has been shown in [32] that the residual block structure can avoid performance degradation caused by the too deep neural network. Therefore, the object

proposal of the robot manipulator is updated from Equation 1 to:

$$\boldsymbol{op}_{\mathcal{M}} := \{\boldsymbol{fmap}_{\boldsymbol{I}_{\mathcal{M}}}, \boldsymbol{fmap}_{\boldsymbol{P}}, \boldsymbol{b}_{\mathcal{M}}, s_{o_{\mathcal{M}}}\}. \quad (3)$$

C. Feature Concatenation

After obtaining $\boldsymbol{op}_i$ from the object detection module and $\boldsymbol{op}_{\mathcal{M}}$ from the position feature extraction network, we are able to recognize the robot manipulator and concatenate its visual and position feature maps as the total feature map $\boldsymbol{fmap}_{\mathcal{M}}$ according to the label results. Until now, the position information of the robot manipulator is integrated into its feature map. Subsequently, we create the feature vectors $\boldsymbol{fv}_i$ from $\boldsymbol{fmap}_{\boldsymbol{I}}$ and $\boldsymbol{fmap}_{\mathcal{M}}$ using an average pooling layer. The updated object proposal $\hat{\boldsymbol{op}}_i$ is given by:

$$\hat{\boldsymbol{op}}_i := \{\boldsymbol{fv}_i, \boldsymbol{b}_i, s_{o_i}\}. \quad (4)$$

D. Relationship Detection Module

The last step of the SRI-Graph generation model is to recognize the predicates between the detected objects $\mathcal{O}$ leveraging the object proposals $\hat{\boldsymbol{op}}_i$. To this end, we adopt the relationship detection network based on Motifs-TDE [16]. It employs the MotifNet [21] as the relationship predictor and Total Direct Effect (TDE) analysis as the post-processing to eliminate the biased relationship prediction. Relationships are created by iteratively integrating object context and relationship context using bidirectional LSTMs on top of $\hat{\boldsymbol{op}}_i$. More details about the architecture of LSTMs and the TDE inference can be found in [16], [21].

Finally, we obtain the detected object labels and the relationships between them in the form of $\langle \textit{subject}, \textit{predicate}, \textit{object} \rangle$. As a result, SRI-Graph is constructed as shown in Figure 4. As suggested in [16], we compute the loss function $\mathcal{L}$ by combining the conventional cross-entropy losses over object classes $\mathcal{L}_o$ and predicate classes $\mathcal{L}_p$: $\mathcal{L} = \mathcal{L}_o + \mathcal{L}_p$. The pseudo-code of the SRI-Graph generation model is given in Algorithm 1.

Algorithm 1: SRI-Graph Generation Model

Input: An RGB image $\boldsymbol{I}$ and the corresponding robot position $\boldsymbol{P}$ at a certain time
Output: SRI-Graph $\mathcal{G}$

```
/* Object Detection Module                    */
Initialize pretrained Faster R-CNN;
for each epoch do
    Update network parameters;
end
Output op_i;
/* Position Feature Extraction                */
for each object o_i do
    if o_i = o_M then
        Extract fmap_P;
        break;
    else
        continue;
    end
end
Output op_M and op_i;
/* Feature Concatenation                      */
for each object o_i do
    if o_i = o_M then
        fmap_M ← concat(fmap_{I_M}, fmap_P);
        fv_i ← average pooling(fmap_M);
    end
    if o_i ≠ o_M then
        fv_i ← average pooling(fmap_I);
    end
end
Output ôp_i;
/* Relationship Detection Module              */
Generate relationships ⟨subject, predicate, object⟩;
Construct G;
```

E. Annotation Tool LabelImg-Rel

It is widely accepted that a quality drop of labeled annotations can lead to a degradation of model performance [33]. For example, the authors in [34] have shown that erroneous annotations deteriorate object detection results. Since the number of relationships between objects that need to be annotated increases quadratically with the number of objects in the scene, the annotation becomes quite time-consuming and is particularly prone to mislabeling.

The most widely used image annotation tools [28], [35] focus only on objects themselves, *e.g.* bounding boxes or masks. Therefore, we develop a tool named LabelImg-Rel to annotate the relationships between objects in a convenient and efficient manner. LabelImg-Rel is based on the open-source image annotation tool LabelImg [28] and allows relationship labeling on the basis of object bounding boxes. As illustrated in Figure 3, the black line from the bounding box of the robot manipulator to the bounding box of the orange container indicates $\langle \textit{robot manipulator}, \textit{right of}, \textit{orange container} \rangle$. We consider the direction of the line to identify *subject* and *object* of the relationship during annotation. Besides, as the annotation format is intended to be applicable to most SGG models, we add the option to export annotations in the same format as the VG dataset [24], which is the most commonly used image dataset for benchmarking.

F. 3DRF-Pos Dataset

Since the existing available image datasets for scene understanding (*e.g.* [24], [25], [36]) do not contain the robot ego information, we recorded a new dataset, named 3DRF-Pos, to perform our study for real-world robot-assisted feeding task. The feeding task is the same as in our previous work [37], where we teleoperate the robot to place the food with a spoon in the containers. The containers are from the YCB real-world object dataset [38]. In addition, we apply an NMPC motion planner [39] to achieve reliable movements of the robot manipulator. To avoid food falling problems during

the various motions of the robot manipulator, the spoon is kept empty during the recording.

3DRF-Pos comprises 760 RGB images with a resolution of 1920×1080 and the corresponding 3D positions of the robot manipulator end-effector in the robot base coordinate system. Each image is associated with 3 objects, 1 robot manipulator, and 12 spatial relationships among them on average. The ground-truth annotation is completed using LabelImg-Rel. As demonstrated in Figure 4, the predefined object classes $\mathcal{C}_o$ are {robot manipulator, red container, orange container, yellow container}, and the predefined predicate classes $\mathcal{C}_\mathcal{P}$ consist of {left of, right of, above, below}.

3DRF-Pos dataset will be announced upon acceptance. This dataset serves as the first use for developing stronger scene understanding models with the help of the robot ego information.

V. Experimental Evaluation

In this section, we evaluate the generated SRI-Graph during a teleoperated 3D real-world feeding task.

A. Experimental Setup

The real experiments are executed on a Intel® Core-i7® CPU with 8 cores. The software is built on Ubuntu® 18.04 LTS and ROS Melodic. We use the Force Dimension® Sigma 7 haptic input/output device to directly control the Kinova® Movo platform for task execution and a ZED 2i stereo camera to record the scene. All SGG models and our SRI-Graph generation model are implemented in PyTorch [40] on Ubuntu® 18.04 LTS using a Nvidia RTX 3090 GPU with 24GB memory.

1) Object Detection Module: Similar as [16], the Faster R-CNN is pretrained on the ImageNet dataset [41]. The batch size is four. The initial learning rate is 1×10^{-3} and decayed by a factor of 10 on the 2k-th and 3k-th iterations. We apply the Stochastic Gradient Descent (SGD) optimizer to update network parameters and thus minimize the loss.

2) Position Feature Extraction: The 10-layers MLP is trained from scratch because the positional information is quite different from the image-based visual information. Hence, unlike other feature extraction works, pretraining on an image dataset does not benefit the initialization of the network parameters.

3) Relationship Detection Module: We adopt the relationship detection network in Motifs-TDE [16] and set the batch size as 8. The learning rate is initialized as 1×10^{-3} and stays static. SGD optimizer is also employed in this module.

4) Metrics: The mean recall@K [42], [43] is selected as the evaluation metric, which represents the percentage of correctly detected relationships among the top K most confident detected relationships. Since most images in our dataset have 12 ground-truth relationships in maximum, we set K = 12 as default.

B. Performance Analysis

To compare the performance of our SRI-Graph with the benchmark approach in [16], we visualize the scene graph results and SRI-Graphs from six input image examples in Figure 4. The selected scenes are very challenging for scene-understanding algorithms using only visual features since the relationships between the robot manipulator and containers are visually hard to determine. We mark the different detected relationships in SRI-Graphs in dark yellow compared with the scene graph results from Motifs-TDE, in which the incorrect relationships are marked in red. Due to the limited space, we selectively show the relevant relationships for comparison.

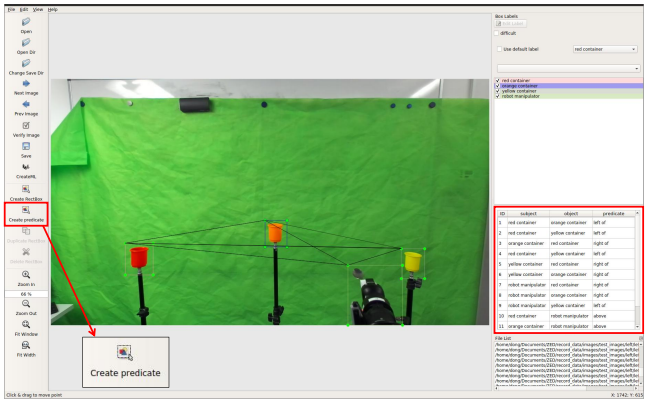

Fig. 3: The interface of LabelImg-Rel. After labeling objects (labels and bounding boxes) in the image, we label the relationships by clicking the subject bounding box, object bounding box, and the button "Create predicate" in the left column. Black lines between bounding boxes indicate the predicates. The detailed relationship information is shown in the right column.

The input images in Figure 4 show the cases when the robot manipulator interacts with the yellow, red, and orange container, respectively. In Figure 4(a), the ground-truth is $\langle$*robot manipulator*, *right of*, *yellow container*$\rangle$. However, the scene graph resulting from Motifs-TDE detects it as $\langle$*robot manipulator*, *left of*, *yellow container*$\rangle$ which is wrong. In Figure 4(c), the Motifs-TDE detects $\langle$*orange container*, *above*, *robot manipulator*$\rangle$, but the ground-truth is $\langle$*orange container*, *below*, *robot manipulator*$\rangle$. In Figure 4(d), the Motifs-TDE mistakenly recognizes that the manipulator is right of the red container, while it is actually left of the red container. In contrast, our method detects the relationships correctly in most scenarios. For Figure 4(b) and (f), our SRI-Graph demonstrates more accurate results as well even though it detects one incorrect relationship.

As shown in Table II, we numerically compare our SRI-Graph with the scene graph generated from [16]. In order to achieve a comprehensive performance analysis, we select three different distributions of the dataset for training/validation/testing, respectively: 1) 110/50/40 images, 2) 190/50/60 images, 3) 398/50/112 images. All images are randomly chosen from the 3DRF-Pos dataset. According to Table II, our SRI-Graph achieves higher mean recall compared to another state-of-the-art model that considers only visual features from RGB images. Note that the main

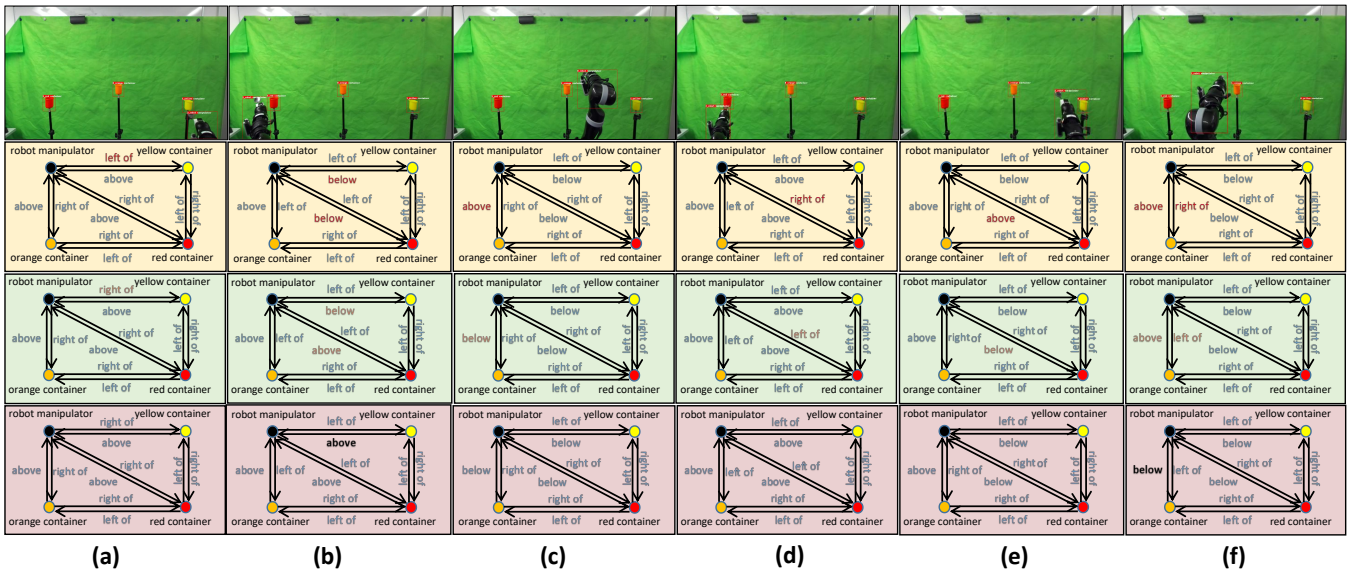

Fig. 4: Comparisons between our model and the state-of-the-art Motifs-TDE model [16]. First row: Input images with bounding boxes. Second row: Results from the Motifs-TDE model. Third row: Results from our SRI-Graphs model. Fourth row: Ground-truth.

TABLE II: Performance Analysis

Model	Dataset Combination	Mean Recall
Motifs-TDE [16]	(1)	75.31%
	(2)	84.11%
	(3)	84.62%
SRI-Graph (Ours)	(1)	**85.14%**
	(2)	**89.50%**
	(3)	**91.04%**

difference between our SRI-Graph model and the Motifs-TDE model is the use of robot 3D positions in our method. This significantly contributes to an improvement of up to 9.83% in detection accuracy.

Compared to the usual scene graph accuracy (normally between ca. 15% - 60 % [16]), all models obtain a noticeable improvement. This is because we use only the 3DRF-Pos dataset containing the robot positions instead of large public image datasets (*e.g.* VG dataset [24], and GQA dataset [25]). Although similar scenarios in the training and test datasets lead to overfitting in our case, the performance enhancement due to the additional robot ego information is clearly observable. In addition, since we consider the manipulation task using the robot where other objects in the scene are usually static, overfitting does not lead to significant deviation in scene understanding for the tested tasks.

In general, we can conclude that our SRI-Graph is able to detect the spatial relationships between the robot manipulator and other objects. Our SRI-Graph achieves robust and accurate results when the relationships are visually uncertain.

Note that the manipulator position features are directly extracted from its 3D positions. Hence these features are relatively sparse compared to the visual features from RGB images. However, our experimental results still show a robust enhancement of the spatial relationship detection between the robot manipulator and objects in the scene.

C. Discussion

Since Motifs-TDE in [16] outperforms other SGG models [17], [19], [20], we show only the comparison results with Motifs-TDE. For a fair comparison, all models are reimplemented using the same object detection module described in Subsection IV-A. As the performance of this module is adequate for our purposes and is not our primary focus, we do not consider other object detection approaches.

VI. Conclusions

In this paper, we have proposed a novel scene-robot interaction graph, namely SRI-Graph, for describing scene understanding information in the robot environment. This scene understanding information includes not only objects and their bounding boxes, but also the relationships between a robot and objects. In our SRI-Graph, we consider the robot ego information *e.g.* 3D positions of the robot manipulator as valuable information in addition to only visual information from RGB images. Our experiments reveal that SRI-Graph outperforms the other conventional SGG models. Since public large-scale image datasets provide only RGB images in daily life scenes and do not contain the robot ego information, we have recorded the 3DRF-Pos dataset during the real-world robot-assisted feeding task. Moreover, a new relationship annotation tool LabelImg-Rel is developed for convenient ground-truth labeling.

Future work will consider two main aspects: To generalize our SRI-Graph to more robot-related tasks, we will investigate how to combine the other robot ego information (*e.g.* robot positions in the environment) with the visual information. Adaptively emphasizing the weight of the robot ego information according to different tasks is also one promising direction.

REFERENCES

[1] G. Zhu, L. Zhang, Y. Jiang, Y. Dang, H. Hou, P. Shen, M. Feng, X. Zhao, Q. Miao, S. A. A. Shah *et al.*, "Scene graph generation: A comprehensive survey," *arXiv preprint arXiv:2201.00443*, 2022.

[2] J. Cheng, L. Wang, J. Wu, X. Hu, G. Jeon, D. Tao, and M. Zhou, "Visual relationship detection: A survey," *IEEE Transactions on Cybernetics*, vol. 52, no. 8, pp. 8453–8466, 2022.

[3] C. Sun, J. M. U. Vianney, Y. Li, L. Chen, L. Li, F.-Y. Wang, A. Khajepour, and D. Cao, "Proximity based automatic data annotation for autonomous driving," *IEEE/CAA Journal of Automatica Sinica*, vol. 7, no. 2, pp. 395–404, 2020.

[4] M. Xiong, X. Xu, D. Yang, and E. Steinbach, "Robust depth estimation in foggy environments combining rgb images and mmwave radar," in *2022 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2022, pp. 34–41.

[5] H. Tian, T. Deng, and H. Yan, "Driving as well as on a sunny day? predicting driver's fixation in rainy weather conditions via a dual-branch visual model," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 7, pp. 1335–1338, 2022.

[6] L. Seenivasan, S. Mitheran, M. Islam, and H. Ren, "Global-reasoned multi-task learning model for surgical scene understanding," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3858–3865, 2022.

[7] G. Humblot-Renaux, L. Marchegiani, T. B. Moeslund, and R. Gade, "Navigation-oriented scene understanding for robotic autonomy: Learning to segment driveability in egocentric images," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2913–2920, 2022.

[8] H. Hu, H. Wang, Z. Liu, and W. Chen, "Domain-invariant similarity activation map contrastive learning for retrieval-based long-term visual localization," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 2, pp. 313–328, 2021.

[9] L. Wellhausen, A. Dosovitskiy, R. Ranftl, K. Walas, C. Cadena, and M. Hutter, "Where should i walk? predicting terrain properties from images via self-supervised learning," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1509–1516, 2019.

[10] G. Kahn, P. Abbeel, and S. Levine, "Badgr: An autonomous self-supervised learning-based navigation system," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1312–1319, 2021.

[11] Z. Wang, F. Mei, X. Xu, and E. Steinbach, "Towards subjective experience prediction for time-delayed teleoperation with haptic data reduction," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2022, pp. 129–134.

[12] K. Chandan, J. Albertson, X. Zhang, X. Zhang, Y. Liu, and S. Zhang, "Learning to guide human attention on mobile telepresence robots with 360 vision," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 5297–5304.

[13] Z. Zeng, A. Röfer, and O. C. Jenkins, "Semantic linking maps for active visual object search," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 1984–1990.

[14] L. Kunze, K. K. Doreswamy, and N. Hawes, "Using qualitative spatial relations for indirect object search," in *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2014, pp. 163–168.

[15] H. Zhang, L. Jin, and C. Ye, "An rgb-d camera based visual positioning system for assistive navigation by a robotic navigation aid," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 8, pp. 1389–1400, 2021.

[16] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased scene graph generation from biased training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3716–3725.

[17] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph r-cnn for scene graph generation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 670–685.

[18] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3668–3678.

[19] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5410–5419.

[20] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene graph generation from objects, phrases and region captions," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1261–1270.

[21] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5831–5840.

[22] S. Sharifzadeh, S. M. Baharlou, M. Berrendorf, R. Koner, and V. Tresp, "Improving visual relation detection using depth maps," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 3597–3604.

[23] S. Y. Gadre, K. Ehsani, S. Song, and R. Mottaghi, "Continuous scene representations for embodied ai," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 849–14 859.

[24] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.

[25] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6700–6709.

[26] S. Amiri, K. Chandan, and S. Zhang, "Reasoning with scene graphs for robot planning under partial observability," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5560–5567, 2022.

[27] S. Li, P. Zheng, Z. Wang, J. Fan, and L. Wang, "Dynamic scene graph for mutual-cognition generation in proactive human-robot collaboration," *Procedia CIRP*, vol. 107, pp. 943–948, 2022.

[28] Tzutalin, "Labelimg," Free Software: MIT License, 2015. [Online]. Available: https://github.com/tzutalin/labelImg

[29] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, and A. G. Hauptmann, "A comprehensive survey of scene graphs: Generation and application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[31] 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.

[32] 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.

[33] B. Pande, K. Padamwar, S. Bhattacharya, S. Roshan, and M. Bhamare, "A review of image annotation tools for object detection," in *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*. IEEE, 2022, pp. 976–982.

[34] K. Alhazmi, W. Alsumari, I. Seppo, L. Podkuiko, and M. Simon, "Effects of annotation quality on model performance," in *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. IEEE, 2021, pp. 063–067.

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

[36] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *European conference on computer vision*. Springer, 2016, pp. 852–869.

[37] E. Babaians, D. Yang, M. Karimi, X. Xu, S. Ayvasik, and E. Steinbach, "Skill-cpd: Real-time skill refinement for shared autonomy in manipulator teleoperation," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 6189–6196.

[38] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The ycb object and model set: Towards common benchmarks for manipulation research," in *2015 international conference on advanced robotics (ICAR)*. IEEE, 2015, pp. 510–517.

[39] S. Hu, E. Babaians, M. Karimi, and E. Steinbach, "Nmpc-mp: Real-time nonlinear model predictive control for safe motion planning in manipulator teleoperation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 8309–8316.

[40] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[41] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large

scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[42] T. Chen, W. Yu, R. Chen, and L. Lin, "Knowledge-embedded routing network for scene graph generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6163–6171.

[43] K. Tang, H. Zhang, B. Wu, W. Luo, and W. Liu, "Learning to compose dynamic tree structures for visual contexts," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6619–6628.