2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Robotic Understanding of Spatial Relationships Using Neural-Logic Learning

Fujian Yan, Dali Wang, and Hongsheng He

***Abstract*—Understanding spatial relations of objects is critical in many robotic applications such as grasping, manipulation, and obstacle avoidance. Humans can simply reason object's spatial relations from a glimpse of a scene based on prior knowledge of spatial constraints. The proposed method enables a robot to comprehend spatial relationships among objects from RGB-D data. This paper proposed a neural-logic learning framework to learn and reason spatial relations from raw data by following logic rules on spatial constraints. The neural-logic network consists of three blocks: grounding block, spatial logic block, and inference block. The grounding block extracts high-level features from the raw sensory data. The spatial logic blocks can predicate fundamental spatial relations by training a neural network with spatial constraints. The inference block can infer complex spatial relations based on the predicated fundamental spatial relations. Simulations and robotic experiments evaluated the performance of the proposed method.**

***Index Terms*—spatial constraints, neural-logic learning, logic rules, cognitive human-robot interaction, deep learning in robotics and automation**

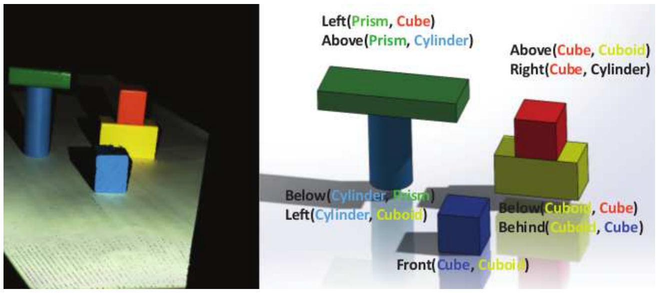


Figure 1: Comprehension of spatial relations by robots with logic-neural network. The Zivid sensor has been used to extract point cloud data. The raw sensory data is used to predict the spatial relations among objects. As a result, robots can understand spatial relationships. The simulated spatial relations that have been generated and illustrated by the Lampix projector to help robot-human interaction.

I. Introduction

Spatial relationships between objects are important compositions of the representation for a physical environment, and it is also a vital description of a scene for human being's daily lives [1]. Understanding spatial relationships are one of the essential abilities for both humans and robots to interact with the physical world [2]. Autonomous robots that worked in the human-centered environment need to understand object's spatial relations [3], [4], besides understanding the semantics of objects [5]–[8], and comprehension of natural language inputs [9], [10]. In the future, we are looking for robots, which can be proficient at tasks such as assembling IKEA furniture, setting up the table, and arranging objects.

The environment of human beings is complicated compared with the well-engineered robots' working space because there are uncertain appearances of variety [11]. There are high chances for robots to interact with objects, which have held uncertain spatial relations because each user may have different preferences [3]. For example, in a collaborative manufacturing scenario, both robots and human workers are working in the same environment. Robots might need to collect parts or tools for workers, and those tools or parts might be named spatial alias instead of using their standard names such as an instruction from the worker may like "give me the part on the right of the workbench." In this situation, robots need to understand fundamental spatial relations and infer complex spatial relations. Whether a robot can deal with these uncertainties is crucial to keep a high-level performance or not. In the meantime, robots should be able to reason complex spatial relations with the fundamental spatial relations that have been learned. So, it is highly needed a well-formed knowledge base, which has every situation of spatial relationships in the real world to aid robots working in a human-centered environment. It is a feasible theoretically method, but it requires too much work to achieve practically [12], [13].

There are some previous researches [14]–[18] in understanding spatial relationships between objects in a scene. Researchers are trying to enable robots to understand spatial relationships between objects so that robots can perform their tasks effectively and efficiently. Some early approaches [19]–[21] define spatial relationships based on 2D images, but the predicted accuracy is limited [22], [23]. Recently, some researches [24], [25] trained data-driven methods based on the point cloud dataset to determine the spatial relationships. Compared with cognition on spatial relationships between objects, there are models [26]–[28] that are hand-

Fujian Yan and Honegsheng He are with Department of Electrical Engineer and Computer Science, Wichita State University, Wichita, KS, 67260, USA fxyan@wichita.edu, hongsheng.he@wichita.edu.

Dali Wang is a Senior R&D Staff and a member of the Artificial Intelligence (AI) team at Oak Ridge National Laboratory (ORNL). wangd@ornl.gov

*Correspondence should be addressed to Hongsheng He, hongsheng.he@wichita.edu.

978-1-7281-6211-9/20/$31.00 ©2020 IEEE

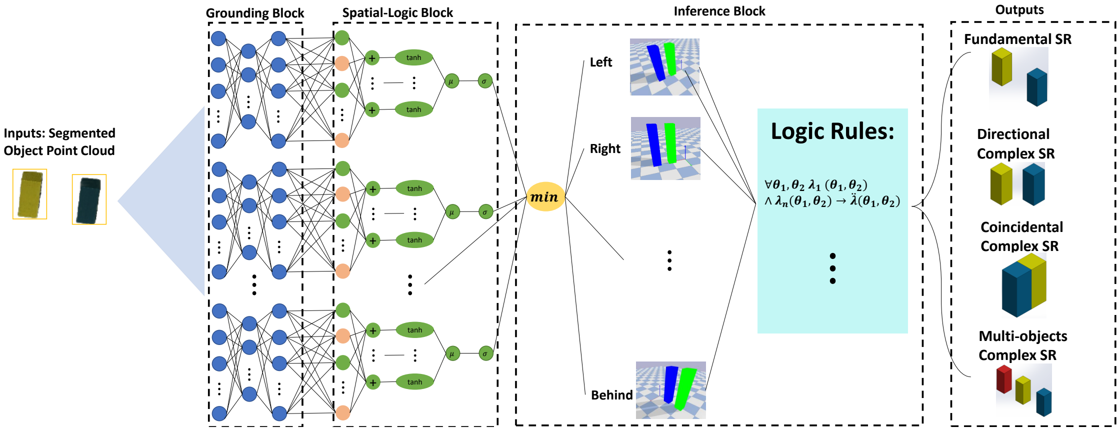


Figure 2: The structure of the proposed neural-logic network. The network has three blocks, which are the grounding block, the spatial logic block, and the inference block.

coding the meanings of spatial relations. Convolution neural network methods are able to learn the fundamental spatial relationships between objects [29]. When the objects in a scene hold complex spatial relationships, there is a huge among of data needed for training. The spatial relationships can be calculated based on the point clouds of objects, but this method normally needs to generate an approximated boundary of each object. It works well when the objects are in simple shapes. The errors can appear when the shapes of objects are complex [30]–[32]. Even though spatial relations can be analyzed based on range sensors [33], as the number of objects increased, the computation will increase exponentially [34]. Besides, some combined spatial relations are not easy to be represented only based on coordination, such as "right front of," "left below," etc.

A spatial comprehension method is needed for fulfilling these challenges [35], [36]. Even though robots can detect spatial relationships due to the development of computer vision and machine learning techniques, conventional perception methods are lack of abilities to adaptive comprehension according to the dynamic environment, and to integrate with prior knowledge [37]. The method proposed in this paper focuses on the autonomous cognition of spatial relationships of objects in the shared environment.

In this paper, we propose a method to comprehend spatial relationships between objects in a scene based on neural-logic learning. Fig. 1 demonstrated the comprehended spatial relation by the proposed system. In contrast to traditional spatial relationships computation [25], [34], we utilize a neural network based logic learning method [38] that can numericize Predicate, Constant, Variable, and logic operators in many-valued first-order logic. In the proposed method, complex spatial relationships can be inferred by the predicated fundamental spatial relationships. We introduce two general types of spatial relations that are useful in assisting robots in a collaborative working environment. These spatial relations are: fundamental spatial relations and complex spatial relations. The complex spatial relations contain three subsets, which are directional complex spatial relations, coincidental complex spatial relations, and multi-object complex spatial relations.

Spatial relationships are symbolic, which is challenging to represent in data forms for numerical training. Logic can represent rich knowledge; however, it is usually complicated and inefficient in representing raw data. We designed a network that learns and reasons spatial relations. The proposed method integrates both the advantages of efficient numerical learning and vibrant representation of logic. The structure of the proposed network is illustrated in Fig. 2, and our method consists of three blocks, which are grounding block, spatial logic block, and inference block. Feed-forward neural networks construct both the grounding block and the spatial logic block. The inference block contains logic rules in form of first order logic to inference complex spatial relations. The network takes point clouds of objects as input. The grounding block of the network is used to extract high-level features from the raw sensory data, and the neural-logic block can predict fundamental spatial relations between objects. The inference block is used to reason complex spatial relations that are useful in assisting robotic comprehension on the scene. The major contribution of this work is an approach that gives robots the ability to learn and infer spatial relationships of objects with acquired sensory data of objects.

II. Neural-logic Network

We propose a neural-logic learning network to comprehend spatial relationships of objects. The input of the network is the pair-wise point clouds of objects in the scene, which are acquired from an RGB-D senor; the output of the network is the spatial relationships among objects.

The logic rules of spatial constraints are learned by using neural network with the pair-wise point clouds of objects in the scene. To extract pair-wise point clouds of objects, we subtract the background of point cloud $\mathrm{P_m} = \{(\mathrm{x_i}, \mathrm{y_i}, \mathrm{z_i}, \mathrm{r_i}, \mathrm{g_i}, \mathrm{b_i})\}, \mathrm{i} = 1 \cdots \mathrm{m}$ for a scene by Faster R-CNN and color-based method. The remaining point clouds of objects $\mathrm{P_n} = \{(\mathrm{x_j}, \mathrm{y_j}, \mathrm{z_j}, \mathrm{r_j}, \mathrm{g_j}, \mathrm{b_j})\}, \mathrm{j} = 1 \cdots \mathrm{n}$, $\mathrm{P_N} \subset \mathrm{P_M}$ are extracted. These extracted pair-wise point clouds are fed into the proposed grounding block to extract high-level features, and then these high-level features are used as inputs for the spatial logic block to recognize fundamental spatial relationships, which are held between objects. The predicated fundamental spatial relationships from the spatial logic block are fed into the inference block to infer complex spatial relations.

The proposed network enables robots to understand the spatial relationships of objects in a scene by learning pre-defined spatial constrains. It can adapt rich knowledge of logic representation as an improvement for predictions, and this enables the network to improve in learning. The advantage of the network is able to map rich symbolic knowledge of these spatial rules to a digital space by using fuzzy logic. The network has three parts, which are the grounding block and the spatial logic block, and the inference block. The structure of the network is illustrated in Fig. 2. The first three layers are grounding layers that can extract high-level features from raw sensory data. The rest layers are used to learn pre-defined spatial rules. In the inference block, complex spatial relationships are inferred from the fundamental spatial relationships that are predicted from the spatial logic block.

A. Grounding Block

Features of point clouds can be extracted manually, but it is complicated and time-consuming to do it that way. Compared with manual feature extraction, it is more efficient to extract high-level features from raw sensory data by using the neural network. These high-level features are beneficial for performing classification tasks. In this paper, the high-level features of raw sensory data are extracted by a fully connected feed-forward neural network.

The grounding block is constructed by one input layer and two fully connected feed-forward layers. There are 48 neurons in the first layer of the grounding block. The second layer has 24 neurons using ReLU as activation function, and the third layer has 48 neurons with ReLU as activation as well. The weights of each layer are initialized using a random normalization method with the mean equal to zero, and the standard deviation equals to one.

The inputs of the grounding block are pair-wise point cloud samples of objects in the scene. The sampling process includes three procedures, which are the background elimination, the point cloud clustering, and the points sampling. In this paper, we used Faster R-CNN [39] model to detect objects in the RGB images that are taken by the same RGB-D sensor. Instead of searching the dominant color range for whole the scene, the ranges that are inside the boundary box are searched. The k-mean clustering algorithm has been used to select the dominant color range. The selected dominant color range in the boundary boxes, which are containing the objects, are used to eliminate the background that is captured by the RGB-D sensor. The structure of the point cloud data of the scene are in the form of $(\mathrm{x}, \mathrm{y}, \mathrm{z}, \mathrm{r}, \mathrm{g}, \mathrm{b})$. The RGB value has been transformed into HSV value to separate the target objects, which are in the pre-defined color range. The second process is using the k-means method to eliminate the segmented point clouds to decrease the effect of the noise that is not eliminated by background subtraction. There are three groups of clusters, and we choose the cluster that has the most point cloud data. The data sampling is the third procedure, and it can sample 8 points in the object's point cloud data. These 8 points are used to represent the object in the scene. A vector with 48 bits represents the pair of object's point cloud data. The first 24 bits are used to represent the object θ_1, and the rest 24 bits are used to represent the object θ_2. The center point is $\mathrm{C} = (\mathrm{x}, \mathrm{y}, \mathrm{z})$, where $\mathrm{x}, \mathrm{y}, \mathrm{and}\, \mathrm{z}$ is the average of each point in the point cloud segment specified in x-axis, y-axis, and z-axis. Then, by comparing each point's Euclidean distance of each point in the point cloud segment with the center point and find the furthest points to generate the bounder points of the region. This procedure ensures that the input of the spatial logic network is same as the output of the grounding block.

B. Spatial Logic Network Block

We designed a fully connected feed-forward neural network, which has four layers to enable robots to understand spatial relationships. The inputs of this network are grounded features, which are extracted from raw sensory data of objects in a scene. We instantiate the features that are extracted from the raw sensory data into variables in the spatial rules to translate spatial rules to numerical space [38]

$$\mathrm{I}(\mathrm{p}) = \sigma(\mu_{\mathrm{P}}^{\mathrm{T}} \tanh(\mathrm{x}^{\mathrm{T}} \mathrm{W}_{\mathrm{P}}^{[1:\mathrm{k}]} \mathrm{x})) \tag{1}$$

where $\mathrm{W}_{\mathrm{P}}^{[1:\mathrm{k}]}$ is a the weight of the network, and it is a tensor in $\mathbb{R}^{\mathrm{mn} \times \mathrm{mn} \times \mathrm{k}}$, σ is the sigmoid function. $\mathrm{I}(\mathrm{p})$ is the output of one network. The x is the extracted features from the grounding block. The μ is the t-norm [40], [41] that has been used to conjunct the neural network. This encoding enables a network to determine the grounding of a clause (e.g. $\mathrm{left}(\theta_1, \theta_2) \rightarrow \mathrm{right}(\theta_2, \theta_1)$) by calculating the literals (e.g. $\mathrm{left}(\theta_1, \theta_2)$) of the clause and then combine those results to calculate the final result. Two networks are combined, so the logic constrains can be mapped to numerical space.

The proposed method can learn the spatial rules through neural network. Fundamental spatial relations include left (L), right (R), above (A), below (B), front (F), behind (Bh), contact (T), and non-contact (NT). The rules which are regulated for spatial logic learning are based on these logic constrains:

$$\forall \theta_1, \theta_2 : \lambda(\theta_1, \theta_2) \rightarrow \lambda(\theta_1, \theta_2) \tag{2}$$

$$\forall \theta_1, \theta_2 : \neg\lambda(\theta_1, \theta_2) \rightarrow \neg\lambda(\theta_1, \theta_2) \tag{3}$$

$$\forall \theta_1, \theta_2 : \lambda(\theta_1, \theta_2) \rightarrow \ \bar{\lambda}\ (\theta_2, \theta_1) \tag{4}$$

$$\forall \theta_1, \theta_2 : \lambda(\theta_1, \theta_2) \rightarrow \neg\lambda(\theta_2, \theta_1) \tag{5}$$

where θ_1, θ_2 denote objects. λ denotes relationships and $\lambda \in \{\mathrm{L}, \mathrm{R}, \mathrm{A}, \mathrm{B}, \mathrm{F}, \mathrm{Bh}, \mathrm{T}, \mathrm{NT}\}$. The rule (2) denotes that if the spatial relations between two objects θ_1 and θ_2 is λ, then the spatial relations between two objects θ_1 and θ_2 is λ. The (3) denotes that if the spatial relationships between two objects θ_1 and θ_2 are not λ, then it infers the spatial relations of these two objects are not λ. In rule (4), the $\{\lambda, \overline{\lambda}\}$ is contrasting spatial relations, such as {left, right}, {front, behind}, {above, below}, and {contact, non-contact}. This rule (4) denotes if the relation between two objects θ_1 and θ_2 is known, then the object θ_2 and the object θ_1 has the contrasting spatial relation such as left and right. The (5) denotes that if the relationship between the object θ_1 and the object θ_2 is known, then the relationship between the object θ_2 and the object θ_1 cannot be the same. These spatial relations rules are used as fundamental constraints for the spatial logic block. The constraints of the rules $\mathrm{C} = \sum_{i}^{n} \mathrm{I}(\lambda)$, where $\mathrm{I}(\lambda)$ is the instantiation of each rule regarding to eight different fundamental spatial relations. Optimization for the learning is to minimize the loss $\mathrm{L} = \underset{\mathrm{C}}{\mathrm{argmin}}(\sum_{\mathrm{j}}^{\mathrm{m}} \mathrm{C})$, where C is the constrain of rules, the m is four since we are using four spatial constrains. The minimum of the loss L has been optimized with an adaptive gradient optimizer with a starting learning rate equal to 0.01. The weights of the spatial logic block are initiated by random with the mean equals to zero, and the standard deviation equals to one. There are totally four layers in spatial logic block, which the first three layers have 48 neurons and use the Tanh function as the activation function. The last layer has eight neurons and uses the sigmoid function as activation function. The spatial logic block is a parallel structure, which each one is used to translate logic atom into numerical space. The fuzzy semantic logic has been used to conjunct each atom in the logic sentence to represent the semantic of each logic sentence.

C. Inference Block

We defined two kinds of spatial relations in this paper. The first type is the fundamental spatial relations between objects, which expresses the directional spatial relationships. The second type of special relations are the complex spatial relations (CSR) among objects, such as "between," that cannot be acquired directly from the spatial logic block. The demonstration of spatial relationships are illustrated in Fig 3.

1) Fundamental Spatial Relationships: Fundamental spatial relationships are directional descriptions of objects in a scene. These fundamental spatial relationships includes: "left (L)," "right (R)," "above (A)," "below (B)," "front (F)," "behind (Bh)," "contact (T)," and "non-contact (NT)."

In addition, these fundamental spatial relationships are transitive. The rule:

$$\forall \theta_1, \theta_2, \exists \theta_3 : \lambda(\theta_1, \theta_2) \wedge \lambda(\theta_2, \theta_3) \rightarrow \lambda(\theta_1, \theta_3) \quad (6)$$

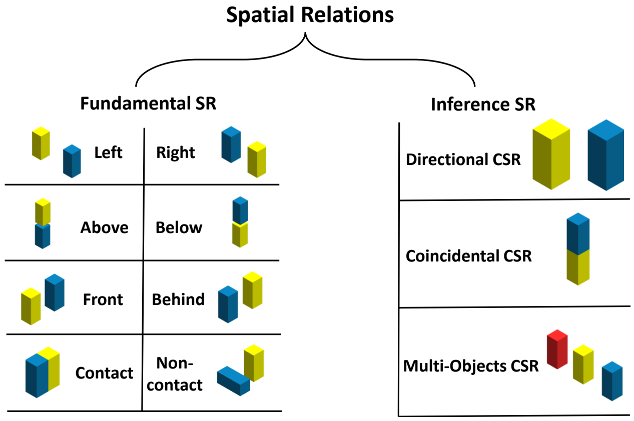


Figure 3: Fundamental and inference spatial relations. These types include fundamental spatial relationships, directional CSR, coincidental CSR, and multi-objects CSR. Objects are yellow, blue, and red cuboid. These objects are generated in SolidWorks.

where θ_1, θ_2, and θ_3 are three objects in the scene, and $\lambda \in \{\mathrm{L}, \mathrm{R}, \mathrm{A}, \mathrm{B}, \mathrm{F}, \mathrm{Bh}\}$. The (6) denotes that if the relation between object θ_1 and object θ_2 is known, and the relation between object θ_2 and object θ_3 is known, then the relation between the object θ_1 and the object θ_3 can be inferred. For an example, $\mathrm{L}(\theta_1, \theta_2) \wedge \mathrm{L}(\theta_1, \theta_3) \rightarrow \mathrm{L}(\theta_1, \theta_3)$, where L stands for the left relationship, and $\theta_1, \theta_2, \theta_3$ are objects. The example shows that if the spatial relation between θ_1 and θ_2 is left, and the spatial relation between θ_2 and θ_3 is also left. Then, the spatial relation between θ_1 and θ_3 can be inferred as left. As shown in the example, each time the spatial logic block predicates the spatial relationships of objects in pairwise and based on these fundamental spatial relationships more complicated spatial relations can be inferred for multi-objects.

2) Complex Spatial Relationships: Besides these fundamental spatial relationships, which are introduced above, there is another type of spatial relationships that is defined as complex spatial relationships (CSR). These complex spatial relationships are inferred based on fundamental relationships. The CSR includes three types directional CSR, coincidental CSR, and multi-objects CSR.

a) Directional CSR: Directional CSR can be used to infer complex directional spatial relationships, the following rule:

$$\forall \theta_1, \theta_2 : \lambda_{\mathrm{L}}(\theta_1, \theta_2) \wedge \cdots \wedge \lambda_{\mathrm{Bh}}(\theta_1, \theta_2) \rightarrow \lambda^{++}(\theta_1, \theta_2) \quad (7)$$

where λ^{++} is the complex directional relations. Both θ_1 and θ_2 are objects. The complex relations λ^{++} cannot be contrasting because the constrain (4) that is trained into the neural network.

b) Coincidental CSR: Another type of CSR is coincidental CSR. The coincidental CSR describes not only the directional spatial relations between two objects, but also whether two objects are touching or not. The rule below:

$$\forall\theta_1,\theta_2 : \lambda(\theta_1,\theta_2)\wedge\tilde{\lambda}(\theta_1,\theta_2)\rightarrow\check{\lambda}(\theta_1,\theta_2) \tag{8}$$

where $\lambda \in \{\mathrm{L},\mathrm{R},\mathrm{A},\mathrm{B},\mathrm{F},\mathrm{Bh}\}$, and $\tilde{\lambda} \in \{\mathrm{T},\mathrm{NT}\}$. $\check{\lambda}$ is the coincidental CSR, which in {LT, RT, FT, BT, O(on), Bn(Beneath)}.

c) Multi-objects CSR: CSR can be extended to three or more objects by comparing the spatial relationships pairwise. For example, $\forall\theta_1,\theta_2,\theta_3$:$\mathrm{L}(\theta_1,\theta_2)\wedge\mathrm{L}(\theta_2,\theta_3)\wedge\mathrm{F}(\theta_2,\theta_3)\rightarrow\mathrm{LF}(\theta_1,\theta_3)$, which denotes that the combine spatial relationship "front-left." Another common spatial relationship among multiple objects is "Between," which is defined by:

$$\forall\theta_1,\theta_2,\theta_3 : \lambda(\theta_1,\theta_2)\wedge\lambda(\theta_2,\theta_3)\rightarrow\lambda_{\mathrm{Bet}}(\theta_2,(\theta_1,\theta_3)) \tag{9}$$

where $\lambda \in \{\mathrm{L},\mathrm{R},\mathrm{A},\mathrm{B},\mathrm{F},\mathrm{Bh}\}$, and λ_{Bet} denotes between.

III. Experiment

A. Experiment Setup

In the experiment, a Zivid depth sensor was used for RGB-D data acquisition. The depth sensor is using stereoscopic technology to depth measurement. It can operate from 0.11m to 10m. The depth resolution is 1280×720, and the depth field of view is 85.2×58. The ROS system for robot controlling run on the machine. The model was trained on Intel Core i7-5930 processors and an NVIDIA TITAN X GPU. A Lampix projector was used to show the feedback from the comprehension of the spatial relations for the robot.

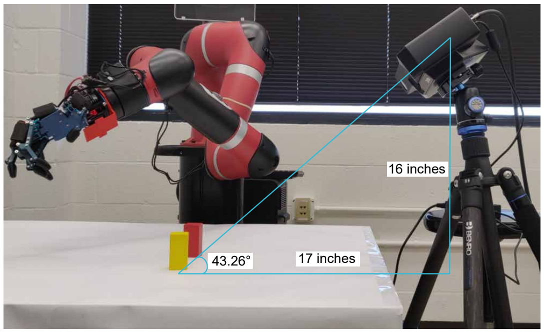


Figure 4: Experiment setup for the robot and the Zivid sensor.

B. Simulation

The prediction accuracy of the proposed method was provided. There were eight spatial relationships were evaluated in this experiment, which were "Left," "Right," "Above," "Below," "Front," "Behind," "Contact," "Non-Contact."

The input data were artificial data that was generated to serve the purpose of simulating point cloud segments of objects. There were totally 20,000 points were generated randomly. These points were generated in the range between 0 to 2. According to these eight spatial relationships, which were mentioned above, these data were divided into eight groups evenly. Due to the data was generated randomly, the numbers of each spatial relationship group were different. We selected 150 pairs of objects for each spatial relationship groups to formulate dataset.

C. Quantities Comparison

In this evaluation, the time efficiency was compared between the neural-logic learning framework with grounding-net and without grounding-net with the same size inputs. We used 2-D artificial data to represent point cloud regions as rectangular with respect to simplicity and time-efficiency. The framework with grounding-net is finished at 200th iterations and reached to 99% accuracy, and the framework without grounding-net took 9900th iterations and reached to 86.89%.

There was another comparison study has been done to compare the proposed network with a four-layer feed-forward neural network. With the same size input data, the neural-logic learning network finished training with 200th iterations and achieved 99% accuracy. The neural network model was not converged within the same training iterations.

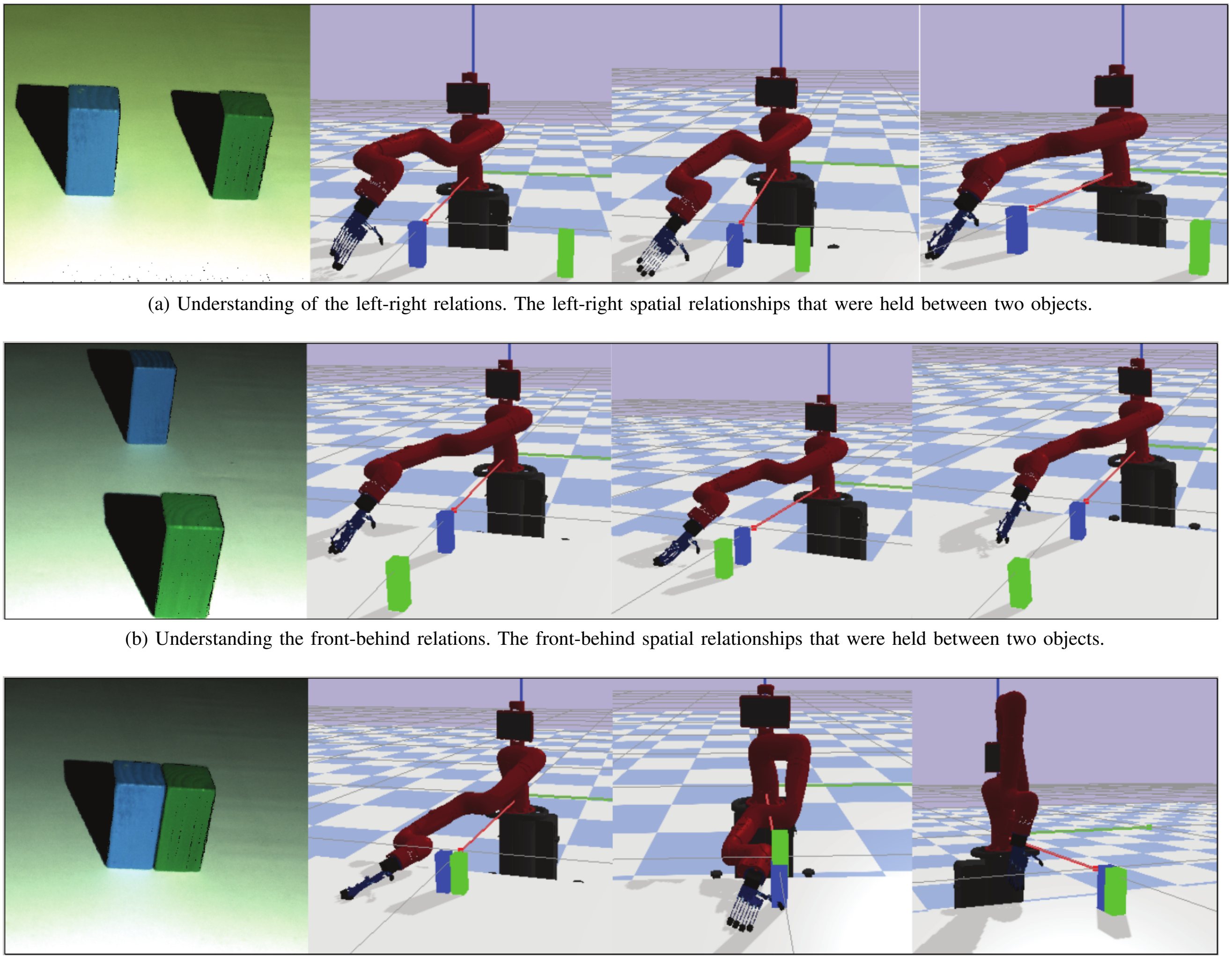
The third comparison study was between other spatial relationships comprehension method [34]. It is a pioneer in the robotic spatial relationship comprehension filed that applied the object point cloud segment and machine learning method. The k-nearest neighbors method has been used to find spatial relationships hold between objects. They used 128 images of working space, and they split 95 of them as training images and 33 of them as testing images. We used 100 pairs of artificial objects as the training set and 25 pairs of artificial objects as the testing set.

The work [25] aimed to recognize two kinds of relationships, which are on and adjacent. For comparison, we compared the above relationship with the on relationship and the adjacent relationship with other relationships. The metric results are shown in Table I. Besides the prediction accuracy of the inside relationship is 98%, the accuracy of other relationships is 99%. The CPN method can predict 40 of 49 images for the on relationship and 39 of 49 for the adjacent relationships. We have done another comparison with another work [25], in that work, the RANSEM method is presented, only left, right relationships in that work is considered.

Table I: Comparison Study of the Prediction Accuracy

Spatial Relationships	**This Paper**	CPN [34]	RANSEM [25]
Left	0.99	0.79	0.86
Right	0.99	0.79	0.89
Above	0.99	0.82	-
Below	0.99	0.79	-
Front	0.99	-	-
Behind	0.98	-	-
Contact	0.99	-	-
Non-Contact	0.98	-	-

The prediction accuracy of the proposed method is considered good compare with previous approaches [24], [34]. Our method is comprehensive because all fundamental spatial



(a) Understanding of the left-right relations. The left-right spatial relationships that were held between two objects.

(b) Understanding the front-behind relations. The front-behind spatial relationships that were held between two objects.

(c) Understanding the contact relations. The contact spatial relationships that were held between two objects.

Figure 5: Demonstrating the simulation results of spatial relations. The Zivid RGB-D sensor had scanned three groups of objects, and the representation of each pair of spatial relations that were comprehended by the robot was demonstrated in the Pybullet simulation environment. The first one on the left was scanned by the Zivid sensor. The rest of the figures were generated in Pybullet environment.

relationships are considered. One limitation for the proposed method is that our method is suitable to resolve precise spatial relationships, on the other hand, some blurred spatial relationships such as “near,” “far,” and “close to” cannot be learned due to that these blurred spatial relationships are not able to define in logic constrains exhaustively.

D. Real-World Scenarios

We have done two experiments that can demonstrate the practical application of the proposed method. In this evaluation, the model was evaluated based on collected real-world data from the Zivid RGB-D sensor. The sample results were shown in Fig. 5. We used several types of toy blocks to test our model. For the practical experiments, we used a green and a blue cuboid to illustrate results. The setup of the Zivid sensor was demonstrated in Fig. 4. The point of view of RGB-D sensor was 23.34 inches away from the objects. The angle between the point of view of RGB-D sensor and objects was approximately 43.26°, which was the ideal angle for obtaining the best quality point cloud of objects.

1) Data Sampling: The point cloud data of the scene was extracted by using the Zivid RGB-D sensor. The reason for using Zivid was the accuracy of the sensor. For this experiment, the input data was in structure of $(\mathrm{x}, \mathrm{y}, \mathrm{z}, \mathrm{r}, \mathrm{g}, \mathrm{b})$. We used a pre-trained Faster R-CNN method to detect the objects in RGB images, and K-mean clusters with a cluster factor equaled 3 to segment the objects from the background of the scene.

2) Demonstration on Comprehended Spatial Relations by Robots: Three samples of results were shown in Fig 5 that can illustrate the robotic comprehension of the scene in spatial aspects. With the given raw sensory data from the scan

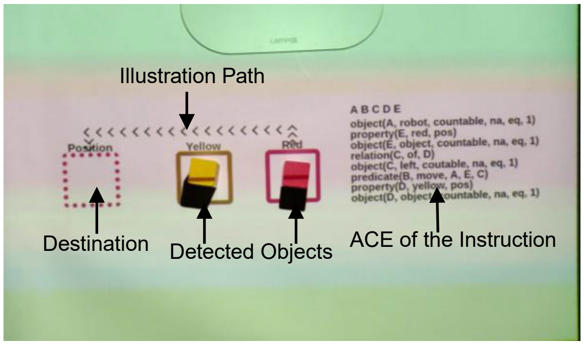


Illustration Path
Position
Yellow
Red
A B C D E
object(A, robot, countable, na, eq, 1)
property(E, red, pos)
object(E, object, countable, na, eq, 1)
relation(C, of, D)
object(C, left, coutable, na, eq, 1)
predicate(B, move, A, E, C)
property(D, yellow, pos)
object(D, object, countable, na, eq, 1)
Destination
Detected Objects
ACE of the Instruction

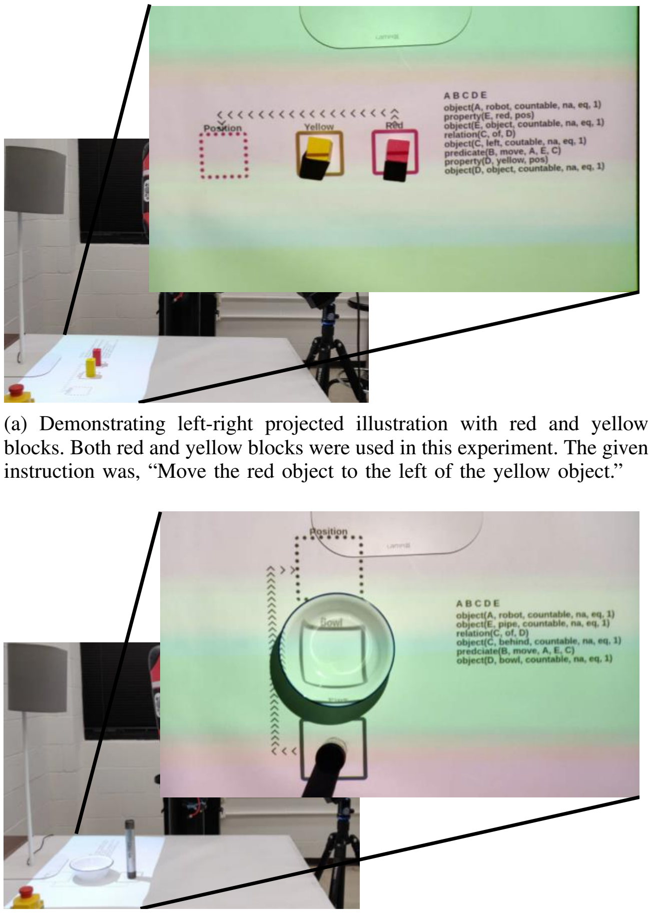
Figure 6: The explanation of the each projected part. In the projected illustration, the detected objects were labeled. A dotted boundary box marked a simulated destination. The result of parsed spatial instruction has been shown. ACE stands for attempto controlled English.

of the scene, robotic comprehension on the corresponding spatial relationships were generated in the Pybullet simulation. Three random demonstrations of the scene for each comprehended spatial relation was generated. The comprehended spatial relations were based on the prediction of the proposed network. As demonstrated in Fig. 5a, two objects in the scene had been scanned by the Zivid sensor. The raw sensory point cloud data was fed into the proposed network to predict the spatial relationships that were held between objects in the scene. The prediction results of the given input were illustrated by generating scenes of the robot's compression, which were based on the predicted results.

3) Demonstration on Spatial Relation Comprehension: The practical experiment has been provided to demonstrate the comprehension of spatial relations to assist human-robot collaboration. A spatial instruction in natural language was given to the robot for each trail of the experiment. A Lampix projector was used to illustrate what the robot has understood from the instruction. In the illustrations, objects were detected by the depth sensor that is embedded in the Lampix projector. A simulated destination had been generated in the illustration. An example of one output from the Lampix projector was shown in 6. Four sample results were shown in Fig 7. The experiment can illustrate that the comprehension of spatial relations can assist robots in understanding instructions. The instruction has been parsed semantically.

IV. Conclusion

In this paper, we have proposed a neural-logic network that enables robots to learn and infer both fundamental spatial relation and complex spatial relations. The proposed model has integrated both the efficiency of data-driven learning and the rich knowledge representation of logic. The simulation can show the predicate accuracy of the proposed model. We have done two practical experiments that can demonstrate the proposed method in assisting the robot in comprehending spatial relations and human-robot interaction.


Position
Yellow
Red
A B C D E
object(A, robot, countable, na, eq, 1)
property(E, red, pos)
object(E, object, countable, na, eq, 1)
relation(C, of, D)
object(C, left, coutable, na, eq, 1)
predicate(B, move, A, E, C)
property(D, yellow, pos)
object(D, object, countable, na, eq, 1)
Position
Bowl
A B C D E
object(A, robot, countable, na, eq, 1)
object(E, pipe, countable, na, eq, 1)
relation(C, of, D)
object(C, behind, countable, na, eq, 1)
predciate(B, move, A, E, C)
object(D, bowl, countable, na, eq, 1)

(a) Demonstrating left-right projected illustration with red and yellow blocks. Both red and yellow blocks were used in this experiment. The given instruction was, "Move the red object to the left of the yellow object."

(b) Demonstrating front-behind projected illustration. Both a pipe and bowl were used in this experiment. The given instruction was, "Move the pipe to the behind of the bowl."

Figure 7: Demonstrating spatial comprehension with an interactive projector. The illustration has shown that the system use the Lampix projector to demonstrate the interaction between human and robots. The Lampix projector can show the comprehended spatial relations that embedded in the instruction. The instruction has been shown as well.

References

[1] R. Smith, M. Self, and P. Cheeseman, "Estimating uncertain spatial relationships in robotics," in *Autonomous robot vehicles*. Springer, 1990, pp. 167–193. 1

[2] S. Lemaignan, R. Ros, L. Mösenlechner, R. Alami, and M. Beetz, "Oro, a knowledge management platform for cognitive architectures in robotics," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2010, pp. 3548–3553. 1

[3] O. Mees, N. Abdo, M. Mazuran, and W. Burgard, "Metric learning for generalizing spatial relations to new objects," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 3175–3182. 1

[4] C. Landsiedel, V. Rieser, M. Walter, and D. Wollherr, "A review of spatial reasoning and interaction for real-world robotics," *Advanced Robotics*, vol. 31, no. 5, pp. 222–242, 2017. 1

[5] F. Yan, D. M. Tran, and H. He, "Robotic understanding of object semantics by referringto a dictionary," *International Journal of Social Robotics*, pp. 1–13, 2020. 1

[6] F. Yan, S. Nannapaneni, and H. He, "Robotic scene understanding by using a dictionary," in *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2019, pp. 895–900. 1

[7] F. Yan, Y. Zhang, and H. He, "Semantics comprehension of entities in dictionary corpora for robot scene understanding," in *International Conference on Social Robotics*. Springer, 2018, pp. 359–368. 1

[8] H. Li, Y. Yihun, and H. He, "Magichand: In-hand perception of object characteristics for dexterous manipulation," in *International Conference on Social Robotics*. Springer, 2018, pp. 523–532. 1

[9] A. B. Rao, H. Li, and H. He, "Object recall from natural-language descriptions for autonomous robotic grasping," in *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2019, pp. 1368–1373. 1

[10] A. B. Rao, K. Krishnan, and H. He, "Learning robotic grasping strategy based on natural-language object descriptions," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 882–887. 1

[11] 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. 1

[12] M. Yu, Y.-J. Liu, Y. Zhang, G. Zhao, C. Yu, and Y. Shi, "Interactions with reconfigurable modular robots enhance spatial reasoning performance," *IEEE Transactions on Cognitive and Developmental Systems*, 2019. 1

[13] M. Alomari, P. Duckworth, D. C. Hogg, and A. G. Cohn, "Learning of object properties, spatial relations, and actions for embodied agents from language and vision," in *2017 AAAI Spring Symposium Series*, 2017. 1

[14] W. Burgard, A. B. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun, "Experiences with an interactive museum tour-guide robot," *Artificial intelligence*, vol. 114, no. 1-2, pp. 3–55, 1999. 1

[15] J. D. Kelleher, G.-J. M. Kruijff, and F. J. Costello, "Proximity in context: an empirically grounded computational model of proximity for processing topological spatial expressions," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2006, pp. 745–752. 1

[16] R. Moratz and T. Tenbrink, "Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations," *Spatial cognition and computation*, vol. 6, no. 1, pp. 63–107, 2006. 1

[17] M. Skubic, D. Perzanowski, S. Blisard, A. Schultz, W. Adams, M. Bugajska, and D. Brock, "Spatial language for human-robot dialogs," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 34, no. 2, pp. 154–167, 2004. 1

[18] D. Wilson, F. Yan, K. Sinha, and H. He, "Robotic understanding of scene contents and spatial constraints," in *International Conference on Social Robotics*. Springer, 2018, pp. 93–102. 1

[19] A. Rosenfeld and A. Kak, "Edge detection," in *Digital picture processing*, 1982, vol. 2, pp. 84–112. 1

[20] R. Krishnapuram, J. M. Keller, and Y. Ma, "Quantitative analysis of properties and spatial relations of fuzzy image regions," *IEEE Transactions on fuzzy systems*, vol. 1, no. 3, pp. 222–233, 1993. 1

[21] K. Miyajima and A. Ralescu, "Spatial organization in 2d segmented images: representation and recognition of primitive spatial relations," *Fuzzy Sets and Systems*, vol. 65, no. 2-3, pp. 225–236, 1994. 1

[22] J. M. Keller and X. Wang, "Learning spatial relationships in computer vision," in *Proceedings of IEEE 5th International Fuzzy Systems*, vol. 1. IEEE, 1996, pp. 118–124. 1

[23] M. Clément, C. Kurtz, and L. Wendling, "Learning spatial relations and shapes for structural object description and scene recognition," *Pattern Recognition*, vol. 84, pp. 197–210, 2018. 1

[24] S. Fichtl, A. McManus, W. Mustafa, D. Kraft, N. Krüger, and F. Guerin, "Learning spatial relationships from 3d vision using histograms," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 501–508. 1, 5

[25] J. Li, D. Meger, and G. Dudek, "Learning to generalize 3d spatial relationships," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 5744–5749. 1, 2, 5

[26] E. Tunstel, "Fuzzy spatial map representation for mobile robot navigation," in *Proceedings of the 1995 ACM symposium on Applied computing*, 1995, pp. 586–589. 1

[27] M. O. Franz, B. Schölkopf, H. A. Mallot, and H. H. Bülthoff, "Learning view graphs for robot navigation," *Autonomous robots*, vol. 5, no. 1, pp. 111–125, 1998. 1

[28] A. Saffiotti, "The uses of fuzzy logic in autonomous robot navigation," *Soft Computing*, vol. 1, no. 4, pp. 180–197, 1997. 1

[29] M. Haldekar, A. Ganesan, and T. Oates, "Identifying spatial relations in images using convolutional neural networks," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 3593–3600. 2

[30] H. A. Sulaiman, M. A. Othman, M. M. Ismail, M. A. M. Said, A. Ramlee, M. H. Misran, A. Bade, and M. H. Abdullah, "Distance computation using axis aligned bounding box (aabb) parallel distribution of dynamic origin point," in *2013 Annual International Conference on Emerging Research Areas and 2013 International Conference on Microelectronics, Communications and Renewable Energy*. IEEE, 2013, pp. 1–6. 2

[31] C. Tu and L. Yu, "Research on collision detection algorithm based on aabb-obb bounding volume," in *2009 First International Workshop on Education Technology and Computer Science*, vol. 1. IEEE, 2009, pp. 331–333. 2

[32] C.-C. Chang and C.-F. Lee, "Relative coordinates oriented symbolic string for spatial relationship retrieval," *Pattern Recognition*, vol. 28, no. 4, pp. 563–570, 1995. 2

[33] H. He, Y. Li, and J. Tan, "Relative motion estimation using visual–inertial optical flow," *Autonomous Robots*, vol. 42, no. 3, pp. 615–629, 2018. 2

[34] B. Rosman and S. Ramamoorthy, "Learning spatial relationships between objects," *The International Journal of Robotics Research*, vol. 30, no. 11, pp. 1328–1342, 2011. 2, 5

[35] E. A. Sisbot and J. H. Connell, "Where is my stuff? an interactive system for spatial relations," *arXiv preprint arXiv:1909.06331*, 2019. 2

[36] P. Hawkins, F. Maire, S. Denman, and M. Baktashmotlagh, "Object graph networks for spatial language grounding," in *2019 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2019, pp. 1–8. 2

[37] G. Marcus, "Deep learning: A critical appraisal," *arXiv preprint arXiv:1801.00631*, 2018. 2

[38] L. Serafini and A. d. Garcez, "Logic tensor networks: Deep learning and logical reasoning from data and knowledge," *arXiv preprint arXiv:1606.04422*, 2016. 2, 3

[39] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99. 3

[40] F. Esteva, L. Godo, and C. Noguera, "First-order t-norm based fuzzy logics with truth-constants: distinguished semantics and completeness properties," 2009. 3

[41] F. Bobillo and U. Straccia, "A fuzzy description logic with product t-norm," in *2007 IEEE International Fuzzy Systems Conference*. IEEE, 2007, pp. 1–6. 3