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

Question Generation for Uncertainty Elimination in Referring Expressions in 3D Environments

Fumiya Matsuzawa[1,2], Yue Qiu[1], Kenji Iwata[1], Hirokatsu Kataoka[1], Yutaka Satoh[1,2]
[1]National Institute of Advanced Industrial Science and Technology (AIST), [2]University of Tsukuba

***Abstract*— We introduce a new task of question generation to eliminate the uncertainty of referring expressions in 3D indoor environments (3D-REQ). Referring to an object using natural language is one of the most common occurrences in daily human conversations; therefore, instructing robots to identify a certain object using natural language could be an essential task in various robotic applications, such as room arrangement. However, human instructions are sometimes uncertain. Existing research on visual grounding using natural language in a 3D environment assumes that the referring expression can uniquely identify the object and does not consider that humans unconsciously give uncertain expressions. When faced with uncertainties, humans ask questions to gain further information. Inspired by the above observation, we propose a method that reduces uncertainty by asking questions when being given an obscure referring expression. The purpose of this method is to predict the positions of all candidate objects that satisfy the referring expressions in a 3D indoor environment and then to ask the appropriate questions to narrow down the target objects from them. To achieve this, we constructed a new 3D-REQ dataset, the input of which is a referring expression with uncertainties in the 3D environment and point clouds, and the output of which is the bounding boxes of all candidate objects satisfying the referring expression and a question to eliminate the uncertainty. To the best of our knowledge, 3D-REQ is the first effort to eliminate the uncertainty of referring expressions for object grounding in 3D environments.**

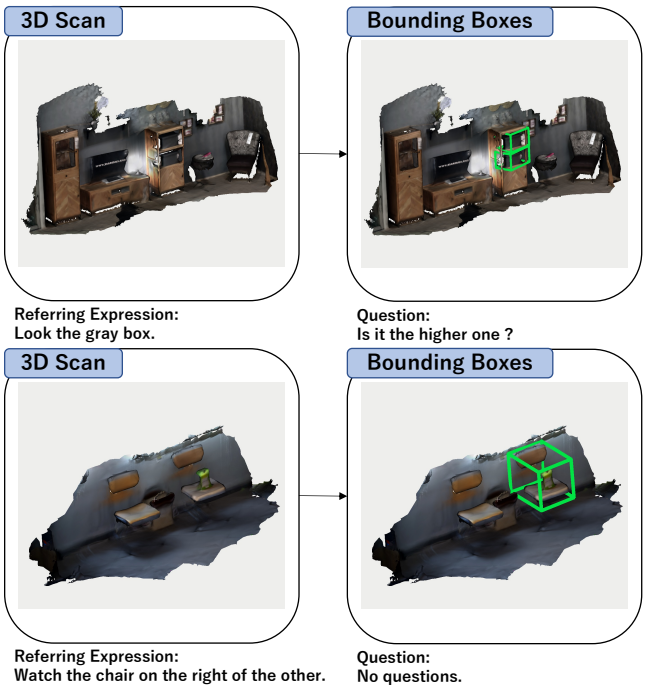

Fig. 1. Illustration of the proposed task. Given a referring expression of an object in a 3D environment (note that unlike existing studies [8], [7], this includes cases where more than one object might be the one being referred to by the expression), our proposed method first detects all objects pointed to by the referring expression and then generates a question to eliminate uncertainties when necessary.

I. INTRODUCTION

In recent years, the accuracy of object detection from two-dimensional (2D) images has greatly improved [1], [2]. More recently, a range of methods [3], [4] have been proposed to detect three-dimensional (3D) objects from point clouds and promising results have been reported in several object detection benchmarks, such as ScanNet [5] and SUN-RGBD [6]. 3D object detection will have a wide range of applications, for example, helping household robots to work efficiently at room organization in a 3D indoor environment.

However, when we want to instruct robots to take some action on a specific object, it is not intuitive to refer to an object by identifying all the details of that object, including its name, attributes, and spatial location, through an interface, such as on a smartphone or computer. It is more natural, intuitive, and effortless for us to refer to an object or give instruction through natural language. To solve this problem, visual grounding, which identifies objects in 3D space using referring expressions (natural language), has been studied[7], [8]. However, these studies assume that the human has spoken all the information necessary to identify the object, when in practice, the robot may take unexpected actions when it receives an uncertain referring expression as input. Therefore, in order to give interpretable instructions to robots, humans would have to carefully observe the scene and pay close attention to ensure that the referring expressions do not contain uncertainty. This is a major problem for practical applications.

When humans are given uncertain instructions, we reduce our uncertainty of the situation by asking questions. Inspired by this, we propose a new method, 3D-REQ, and dataset to eliminate uncertainty in referring expressions in 3D environments. In detail, as shown in Figure 1, our dataset consists of a 3D Scan, a referring expression describing a specific object, and a question for eliminating the uncertainty within the expression if necessary. This dataset contains questions that can be used to properly sort out and reduce uncertainty when multiple candidate objects are suggested by the referring expression.

We also proposed a 3D object detector-based method which detects all objects pointed to by the expression and

979-8-3503-2365-8/23/$31.00 ©2023 IEEE

generates a question to eliminate uncertainty. Using this method, even if a human unintentionally gives an indeterminate instruction to the robot, the robot can appropriately ask the human for the details of the instruction and can communicate with the human in a way that is more similar to daily life.

II. RELATED WORK

A. 2D Referring Expression for Visual Grounding

In recent years, there has been a dizzying growth in research across the fields of image recognition and natural language processing. Visual grounding is one of the representative tasks which integrates image recognition with natural language processing, tasked to localize image regions from referring expressions in natural language. Hu et al. [9] proposed a method for scoring candidate regions based on CNN, while Deng et al. [10] proposed TransVG, which is an end-to-end object grounding method based on the Transformer structure. More recently, 3D visual grounding with referring expression in 3D environments has also been studied [7], [8].

B. Eliminating Ambiguity in Robotics

Natural language is often accompanied by ambiguity. Hatori et al. [11] proposed a method for reducing ambiguity in human instructions for the robot picking task. The target candidate objects are drawn on the display, and the operator is asked to provide further explanations. In their method, an ambiguous situation is defined as a situation where multiple objects have high levels of bounding box confidence. Although Hatori et al. tried to recognize the object ambiguities in robot picking, their proposed model does not output additional information to distinguish those objects. In contrast, in our study, the model learns to recognize the ambiguity contained in referring expressions and outputs question sentences to mitigate them.

C. 3D Object Detection

In recent years, many object detection methods for 3D point clouds have been studied. In particular, object detection based on PointNet [12] and PointNet++ [13] has been widely used. Qi et al. proposed VoteNet [3], a CNN-based 3D object detection method that uses Hough Voting [14] to achieve 3D object detection independent of 2D object detectors. Misra et al. proposed 3DETR [4], a transformer-based end-to-end method which outperforms VoteNet in ScanNetv2 [5], one of the most widely used large-scale datasets of 3D environments for 3D object detection. Wald et al. proposed 3RScan [15], which is a large-scale 3D indoor environment dataset. 3RScan contains 1,482 scans taken at different time steps for 478 indoor environments. 3DSSG [16] is a dataset constructed based on 3RScan for scene graph generation. We created our dataset based on 3DSSG because the 3D scene graphs are annotated for each scan.

D. 3D Vision and Language

While research across image and language has been widely discussed in recent years, 3D vision and language is a relatively new research topic. Here we introduce several representative tasks related to 3D vision and language. VLN [17] is a task of interpreting natural language instructions in an unknown indoor environment and navigating to a goal point. Along with VLN task, Anderson et al. also proposed the Room-2-Room dataset, which contains instructions for traversing 21,567 rooms based on the Matterport3D [18], which is a large 3D indoor environment dataset. It also predicts the robot's viewpoint movement and other actions by using a sequence-to-sequence neural network with the robot's first-person viewpoint as input. Finally, as a couple studies on understanding object features and spatial information from 3D point clouds using language, Angel et al. proposed Scan2Cap[19], which generates bounding boxes and their corresponding captions in a 3D environments, and Azuma et al. proposed ScanQA [20], which generates bounding boxes of objects in a 3D environment and their corresponding answers to given questions.

In this study, we focus on the task of 3D visual grounding, which locates objects in 3D environments by using natural language-based referring expressions. There are some benchmark datasets for the 3D visual grounding task: ScanRefer [8], which contains free referring expressions, Nr3D [7], which contains interactive sentences, and Sr3D [7], which is a template-based referring expression dataset. The referring expressions contained in these datasets can identify a single object in the 3D environment, and thus do not assume that the input sentence is uncertain. In our proposed task, when there is uncertainty in the referring expressions in the 3D visual grounding setting, the model outputs all the bounding boxes of the candidate objects and generates question sentences to appropriately narrow down the target object from the candidate objects.

III. DATASET

Datasets for the task of visual grounding by referring expressions in a 3D environment can be divided into two types: collected from human utterances and generated from 3D information in a rule-based method. The datasets containing human utterances are the ScanRefer and the Nr3D. The rule-based dataset is the Sr3D which consists of referring expressions describing the spatial relationships between two objects. Both types of datasets contain only referring expressions that uniquely identify objects and do not contain uncertainty referring expressions. In addition, the output is only bounding box information of the object identified by the referring expressions.

As mentioned above, existing research does not deal with uncertain representations. However, in the real world, there are many referring expressions that include uncertainty, and it is necessary to consider this when making them applicable. Therefore, we attempt to eliminate the uncertainty of referring expressions by constructing a dataset that includes appropriate questions in natural language, as human usually

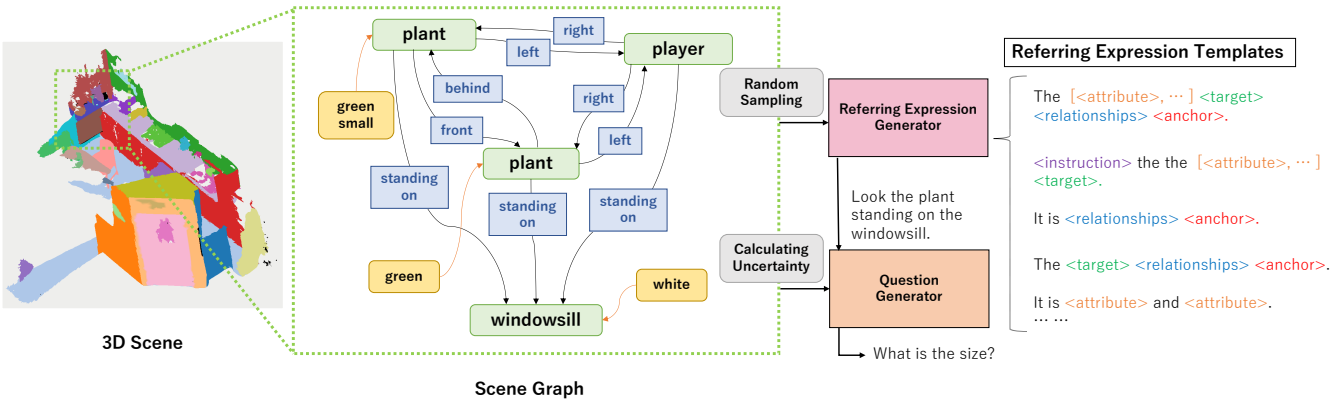

Fig. 2. Dataset generation process of the proposed dataset 3D-REQ.

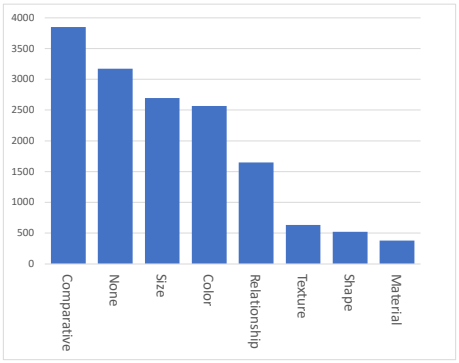

Fig. 3. Question types distribution of 3D-REQ.

use. The dataset we propose contains referring expressions in each 3D scene, bounding boxes for all candidate objects that satisfy the referring expression, and questions asking about the feature or relationship that is causing the uncertainty, e.g., "What color is it?".

A. 3DSSG

Our dataset is created based on 3DSSG following a rule-based generation process. Here, we give some details of the dataset. 3DSSG is a large-scale indoor 3D dataset, and consists of 3D meshes and is densely annotated with 3D scene graphs. In detail, 3DSSG is annotated with two main types of attributes and three types of relationships. The attributes are divided into static attributes such as color, size, shape, and material, and dynamic attributes such as "on/off" and "full/empty". In our proposed dataset, only static features are used. Relationships are divided into supportive relationships such as "standing on" and "lying on", proximity relationships such as "close by" and "next to", and comparative relationships such as "bigger than" and "darker than". An object that has some relationship with the target object is called an anchor. Considering validity for inclusion in the referring expression, classes with three or more objects in the same scene are removed from the anchor. Also, if the target and anchor are objects of the same class, bidirectional relationships such as "close by" are removed.

B. Dataset Generation

The input to this dataset is referring expressions containing point clouds and the possibility that an object cannot be uniquely identified, and the outputs are bounding boxes for all candidate objects and a question sentence to eliminate the uncertainty. Pairs of referring expressions and questions are generated by the following steps, as shown in Figure 2. First, target objects are selected from the objects annotated by the 3DSSG dataset and attributes and relationships are randomly sampled to generate referring expression sentences. The generated referring expression sentences do not necessarily uniquely identify a single object in the scene, and there may be objects other than the target object with identical attributes and relationships. These objects are defined as uncertain objects, and the number of them is an uncertain value. For these uncertainties, among the attributes that are not used in the referring expression, a question statement is selected about the feature that would have the lowest expected value of indeterminacy after the question, rather than a question statement that considers a specific feature of the target object. The expected value of uncertainty $E(U)$ is expressed for the several instances (e.g., red, blue, round, triangular) in the selected attributes (e.g., color, shape) as the following equation, where p_k is the probability that the attribute instance is selected and x_k is the uncertainty when the attribute instance is selected.

$$E(U) = \sum_k x_k \times p_k \tag{1}$$

If the target corresponding to the selected question does not have any attributes, or if the question for any question item does not eliminate the uncertainty, that question text and referring expression pair is not added to the dataset. Otherwise, it is added to the dataset, the attributes of the target corresponding to the selected question are added to the referring expression, and the process of selecting a new question is executed.

C. dataset statistics

We created 15,469 referring expression sentences and questions for each of 3DSSG's 1,122 scans, including 7,361

TABLE I

3D GROUNDING DATASET COMPARISONS.

Datasets	Tasks	Scans	Referring Expression	Uncertainty	Collection	Environment
ScanRefer [8]	Grounding	800	51,583	No	Human	ScanNet
Nr3D [7]	Grounding	707	41,503	No	Human	ScanNet
Sr3D [7]	Grounding	707	83,572	No	Template	ScanNet
Ours	Grounding, Question Generation	1,122	15,469	Containing	Template	3RScan

Fig. 4. Word clouds of all referring expression sentences in 3D-REQ.

TABLE II

TRAINING, VALIDATION, AND TEST SPLITS OF 3D-REQ.

	Train	Validation	Test
Scene	967	117	38
Referring Expression	13,112	1,822	535

unique referring expression lists. The question types in 3D-REQ are as follows and are shown in Figure 3: a comparative question with a specific comparison item, such as "Is it the higher one?"; a relationship question question about location, such as "Where is it?"; and "None" when there is no uncertainty in the referring expression. See Figure 1 for two examples of the 3D-REQ. We also show the word clouds of 3D-REQ in Figure 4, the dataset comparison in Table I, and the dataset splits in Table II.

IV. PROPOSED METHOD

To address the task of generating questions to eliminate the uncertainty of 3D referring expressions, as shown in Figure 5, we propose an end-to-end method that takes 3D point clouds and sentences referring to objects in the 3D environments as inputs and outputs bounding boxes of all objects that satisfy the sentences and question sentences to eliminate their uncertainty.

A. Input Data

One model input item includes a 3D point cloud and a referring expression sentence. The point cloud data contain 20,000 points. Since features such as color data and texture are important in the referring expressions, each point is represented by a total of 6 dimensions, the 3D coordinates XYZ and the RGB values.

B. 3D and Language Feature Extraction

Our method uses PointNet++ as the backbone for point feature extraction from point clouds, similar to VoteNet and 3DETR. Specifically, for an input point cloud of $N \times 6$ dimensions, PointNet++ hierarchically extracts point features from the local region to the degenerate region to obtain $M \times (3 + C)$-dimensional point features. For each point downsampled to M, 3D coordinates and feature dimension C are obtained. Furthermore, the referring representation is converted to a $S \times T$-dimensional feature vector at the embedding layer. Here, S is the maximum number of words.

C. Transformer Encoder and Decoder

The obtained point features are encoded by the self-attention operation of the transformer encoder layer, and the output and the features of the embedded referring expression are used as the inputs for cross-attention. The encoded linguistic features are input to the questioner, which is described below, and the point features are input to the transformer decoder layer to obtain $K \times (3 + C)$-dimensional domain features.

D. Detection Head and Questioner

Given a $K \times (3 + C)$-dimensional region feature, we use a detection head to identify candidate objects in the reference representation and a questioner to generate a question sentence to properly classify the candidate objects. Specifically, the detection head performs a 3D bounding box regression to predict whether each region is a candidate object for the reference representation. Here, referring to 3DETR, the bounding box information is predicted as $\hat{b}$, and $\hat{b} = [\hat{c}, \hat{d}]$. Here $\hat{c}, \hat{d} \in [0, 1]^3$ denote the center and size of the bounding box, respectively. The questioner employs a standard transformer decoder and uses self-attention and feed-forward networks for processing sentence features.

E. Loss Function

This method simultaneously performs candidate object detection of reference expressions and generation of interrogative sentences to reduce uncertainty. For candidate object detection, the dichotomous technique set matching used in DETR and 3DETR is used.

$$Loss_{det} = \lambda_c \|\hat{c} - c\|_1 + \lambda_d \|\hat{d} - d\|_1 - \lambda_s s_c^T log \hat{s}_c \quad (2)$$

We use standard cross-entropy loss $Loss_q$ for question generation and the overall loss function is defined as follows.

$$Loss = \lambda_1 Loss_{det} + \lambda_2 Loss_q \quad (3)$$

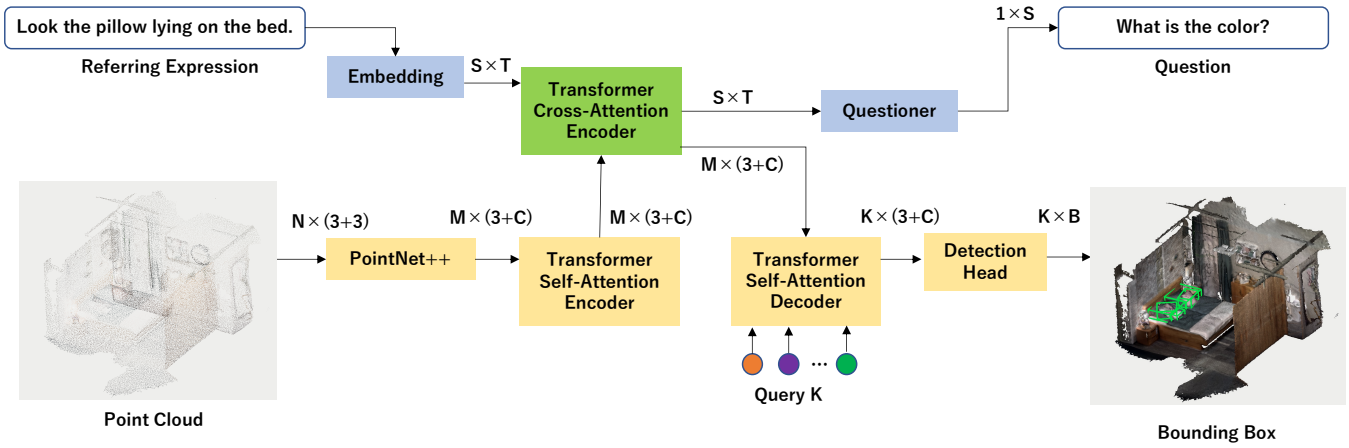

Fig. 5. Illustration of the proposed method.

TABLE III

EVALUATION OF DIFFERENT MODEL ABLATIONS ON QUESTION GENERATION AND OBJECT DETECTION APPLIED TO THE PROPOSED DATASET.

Input	Query	Question generation				Referring expression	
		BLEU	CIDER	METEOR	ROUGE	mAP(0.25)	mAR(0.25)
Geo	32	38.2	228.1	25.9	67.5	11.2	34.4
Geo	64	36.4	197.2	25.7	65.7	**11.8**	42.6
Geo	128	28.4	163.3	21.5	43.9	7.4	43.7
Geo+RGB	32	35.5	218.5	24.4	65.7	9.9	35.4
Geo+RGB	64	**41.9**	**239.9**	**26.9**	**69.4**	10.6	**49.4**
Geo+RGB	128	26.0	130.5	20.3	49.9	5.6	38.7

TABLE IV

PER-OBJECT-CLASS DETECTION EVALUATION OF GEO (QUERY NUMBER 64).

	wall	pillow	chair	shelf	box	table	picture	plant	cabinet	door
mAP(0.25)	25.7	12.5	**42.8**	0.4	0.03	3.9	2.6	1.1	7.9	21.0
mAR(0.25)	46.7	34.1	**70.1**	14.7	16.0	28.5	15.4	17.7	43.3	52.4

V. EXPERIMENTS

A. Experimental Setup

We used our proposed dataset 3D-REQ for model performance evaluation with the training and test split shown in Table II. The goal was to find the bounding boxes of all candidate objects that satisfy the referring expressions and generated questions to mitigate uncertainties.

Implementation Details. We evaluated the proposed model introduced in the previous section. We implemented the questioner with a transformer decoder structure. We set the number of layers and heads to two for the transformer cross-attention encoder, the self-attention encoder, the self-attention decoder, and the transformer questioner. During training, losses in equations (2) and (3) were set as $\lambda_c = 1$, $\lambda_d = 1$, $\lambda_s = 0.1$, $\lambda_1 = 1$, and $\lambda_2 = 0.5$. The size of the input point cloud was set to 20,000 points for all experiments. The learning rate was set to 0.0001, and all model ablations were trained for 40 epochs for all experiments.

Ablations. We conducted ablation experiments on the model input and the query number. In detail, we compared model performance with 3D point clouds, which only contain the 3D coordinates of points, and 6D point clouds, which consist of points having both XYZ and RGB values. We experimented with query numbers of 32, 64, and 128.

Evaluation Metrics. In 3D-REQ, one of the model's goals is to detect all candidate objects corresponding to the referring expression in the form of bounding boxes. Here, we use the conventional evaluation metrics mAR (mean average recall) and mAP (mean average precision) in 3D object detection, where we compute the accuracy of detected bounding boxes that have IoU (intersection over union) greater than 0.25 with the ground truth bounding boxes. Along with bounding boxes, the models also generate a question to reduce the uncertainty. To evaluate the efficiency of generated questions, we used four evaluation metrics

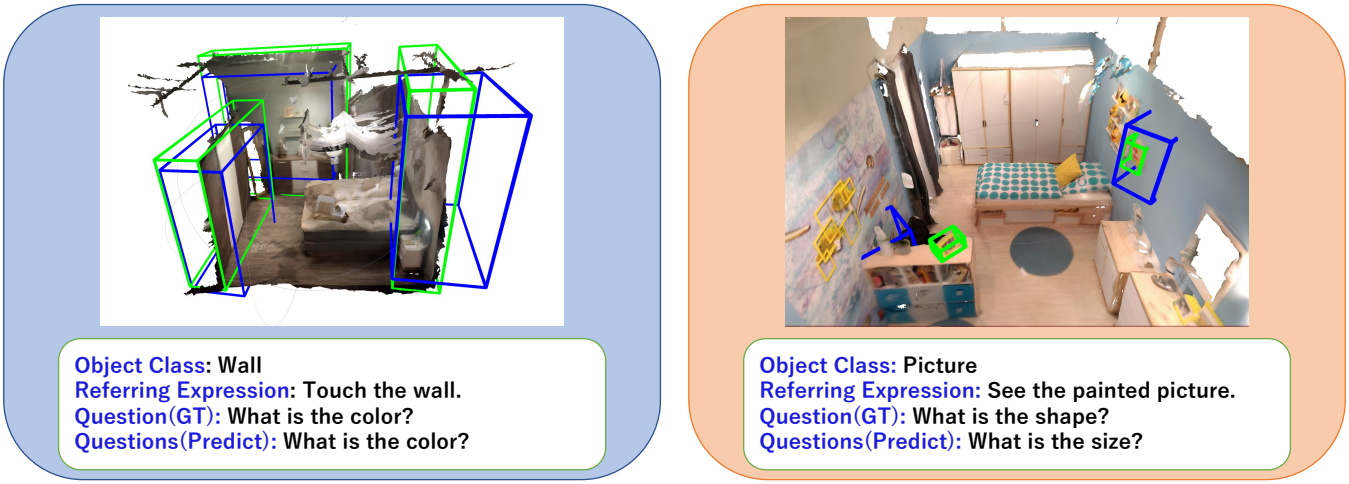

Fig. 6. Qualitative results of Geo+RGB (query number 64). Predicted boxes are marked with blue and ground truth boxes are marked with green. We show examples where our method produced good predictions of the boxes and question (blue block), as well as a failure case (orange block).

which are widely adopted in image captioning tasks, namely, BLEU [21], CIDER [22], METEOR [23], and ROUGE [24]. These four metrics evaluate the similarity between the generated sentences and the ground truth sentences.

B. Quantitative Results

First, Table III shows the quantitative results. The results of the question generation, which is the main objective of this study, are listed in the center four columns of Table III. For both Geo and Geo+RGB, the highest accuracies were obtained when the query number was 64. Here, we found that increasing the number of query did not necessarily improve performance. We also found that the Geo+RGB (query number 64) obtained the highest performance for all four evaluation metrics for question generation. Since the proposed dataset 3D-REQ includes data that can distinguish objects by color, it can be assumed that the presence of color information is beneficial for question generation.

The right side of Table III shows the results of 3D object detection. Here, Geo, which uses only geometric information, achieved higher accuracy than Geo+RGB, which uses color information in addition to geometric information. Although the proposed dataset 3D-REQ includes questions about color, it is difficult to improve the accuracy by simply adding RGB values because the detector used is designed to detect geometric information. We plan to study the effective use of RGB values in future work. Similar to the results for question generation, the highest accuracies were obtained when the number of queries was 64 for both Geo and Geo+RGB.

The detection accuracy per object for the model Geo (query number 64) is listed in Table IV. Here, the accuracies for relatively large objects, such as wall, chair, door, were relatively high, while the accuracies for small objects, such as box, picture, plant, tended to be low. Improvement of accuracies for small objects will be considered in future work.

C. Qualitative Results

Two example results of Geo+RGB (query number 64) are shown in Figure 6. First, on the left side, the referring expression is “Touch the wall”. Here, there are a total of three walls, and our proposed method was able to detect all three walls. The color of the three walls is different, and therefore the color information can be used to distinguish the walls. Here, our proposed method was also able to generate the question “What is the color?”. This example shows that our proposed method is able to detect the objects pointed to by the referring expression and generate sentences to distinguish them. Another example is shown on the right. Here, the referring expression is “See the painted picture”. Here, there are two small paintings. As shown in the green bounding box, each of them has a different shape. In this example, our method did not detect the painting correctly and could not generate a sentence that distinguishes between the two paintings. Currently, our proposed method still has room for improvement for small objects, and we will address this issue in future work.

VI. CONCLUSIONS

In this paper, we propose a new task and dataset aiming at eliminating the uncertainty of referring expressions in 3D indoor environments by generating questions. Finding certain objects specified by natural language sentences is an important task in robotic applications. However, human instructions sometimes contain uncertainties. Existing studies on 3D visual grounding have assumed that the referring expression can uniquely identify the object, but this neglects the fact that humans can unconsciously give uncertain expressions. We address this issue by proposing a framework that locates 3D objects specified by expressions and generates questions for distinguishing those objects. Our experimental results show that our proposed method can generate questions for eliminating ambiguity, and our proposed dataset can be used as a benchmark dataset for future research in uncertainty elimination in 3D visual grounding.

REFERENCES

[1] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[2] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[3] Charles R. Qi, Or Litany, Kaiming He, and Leonidas J. Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[4] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2906–2917, October 2021.

[5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[6] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[7] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. ReferIt3D: Neural listeners for fine-grained 3d object identification in real-world scenes. In *16th European Conference on Computer Vision (ECCV)*, 2020.

[8] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. *16th European Conference on Computer Vision (ECCV)*, 2020.

[9] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4555–4564, 2016.

[10] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1769–1779, October 2021.

[11] Jun Hatori, Yuta Kikuchi, Sosuke Kobayashi, Kuniyuki Takahashi, Yuta Tsuboi, Yuya Unno, Wilson Ko, and Jethro Tan. Interactively picking real-world objects with unconstrained spoken language instructions. In *Proceedings of International Conference on Robotics and Automation*, 2018.

[12] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[13] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[14] Bastian Leibe, Ales Leonardis, and Bernt Schiele. Combined object categorization and segmentation with an implicit shape model. In *Workshop on statistical learning in computer vision, ECCV*, volume 2, page 7, 2004.

[15] Nassir Navab Federico Tombari Matthias Niessner Johanna Wald, Armen Avetisyan. Rio: 3d object instance re-localization in changing indoor environments. 2019.

[16] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[17] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[18] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

[19] Zhenyu Chen, Ali Gholami, Matthias Niessner, and Angel X. Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3193–3203, June 2021.

[20] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19129–19139, June 2022.

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics*, pages 311–318, 2002.

[22] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[23] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[24] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.