OCID-Ref: A 3D Robotic Dataset with Embodied Language for Clutter Scene Grounding

Ke-Jyun Wang∗ Yun-Hsuan Liu∗ Hung-Ting Su Jen-Wei Wang
Yu-Siang Wang Winston H. Hsu Wen-Chin Chen
National Taiwan University

Abstract
To effectively apply robots in working environments and assist humans, it is essential to develop and evaluate how visual grounding (VG) can affect machine performance on occluded objects. However, current VG works are limited in working environments, such as offices and warehouses, where objects are usually occluded due to space utilization issues. In our work, we propose a novel OCID-Ref dataset featuring an referring expression segmentation task with referring expressions of occluded objects. OCID-Ref consists of 305,694 referring expressions from 2,300 scenes with providing RGB image and point cloud inputs. To resolve challenging occlusion issues, we argue that it’s crucial to take advantage of both 2D and 3D signals to resolve challenging occlusion issues. Our experimental results demonstrate the effectiveness of aggregating 2D and 3D signals but referring to occluded objects still remains challenging for the modern visual grounding systems. OCID-Ref is publicly available at https://github.com/lluma/OCID-Ref

Figure 1: A hard case where visual grounding (VG) network fails to predict the occluded object in clutter scene. Our dataset provides more such cases than other datasets, which are commonly seen in the working spaces, like offices and warehouses.

1 Introduction
Visual grounding (VG), which aims to locate the object according to a structured language query, is a crucial task in natural language processing (NLP), computer vision (CV), and robotics. Recent VG studies most focus on web-crawled images such as (Kazemzadeh et al., 2014; Krishna et al., 2016; Mao et al., 2016; Yu et al., 2016). However, VG for human-robot interaction (HRI) is less explored. Most of the images in existing VG datasets are people and daily necessities, e.g., RefCOCO contains mainly persons, cars, and cats, which are separated and therefore easier to detect. Nevertheless, working spaces such as offices or warehouses, where robots are usually applied to assist works, are usually crowded, and objects are overlapped with each other to utilize space better. Therefore, objects in working environments are often occluded and hard to detect.

Previous work (Ralph and Moussa, 2005) suggested that a system that uses language for human-computer interaction can help non-professionals instruct robots to complete technical work and collaborate. Recent research pointed out that VG plays an important role in HRI. (Shridhar and Hsu, 2018) utilized VG to resolve ambiguity in grasping tasks. (Matuszek) studied how the robot learns about objects and tasks in an environment via nature language queries. Therefore, an apparent language instruction/good referring (grounding) expression is pivotal in human-robot interaction and improves the communication between non-expert humans and robots.

Some efforts have been made to collect VG datasets. RefCOCO (Yu et al., 2016) and Cops-
Ref (Chen et al., 2020a) utilize web-crawled images and manually label language expressions. A limitation is that images alone do not provide precise position cues, which are essential for various robotic tasks such as grasping. A recent work, Sun-Spot (Mauceri et al., 2019), utilizes a depth channel for object detection and referring expression segmentation tasks. Another existing dataset, ScanRefer (Chen et al., 2019), uses more accurate multi-view point clouds for 3D signals. However, both Sun-Spot and ScanRefer do not address occlusion issues, which is ordinary in working spaces and more challenging due to more compositions of shapes of each object. As shown in Figure 1.a, when an object *(the red plastic bag)* is blocked in an occluded environment, the shape of the object could be deformed and increase VG difficulty.

Observing this, we propose a novel OCID-Ref dataset with 3 key features: (1) Both RGB image and pointcloud for each scene to provide multimodal signals for learning system development and (2) Scenes with higher clutter level, as shown in Table 1:b, to evaluate the model capability for resolving challenging occlusion issues. To the best of our knowledge, OCID-Ref is the only existing dataset supporting above features, and therefore allows VG task in grasping scenario.

Experimental results demonstrate that occluded scenes are more challenging to modern VG baselines. We observe 27% to 34% performance drops on referring expression segmentation tasks. Also, utilizing 3D information continually improves performance across all clutter levels. Furthermore, fusing 2D and 3D features reach the best performance on all clutter levels. We suggest that OCID-Ref dataset could pave a new path for VG research in HRI and benefit the research community and application developments.

2 Dataset and Task

To open up a new way for VG research in HRI, we collect a novel OCID-Ref dataset by the following steps: (1) We leverage a robotic object cluttered indoor dataset, OCID (Suchi et al., 2019) which consists of complex clutter-level scenes with richful 3D point cloud data and the point-wise instance labels for each occluded objects. (2) We manually annotate fine-grained attributes and relations such as color, shape, size relation or spatial relation. (3) We generate referring expressions based on annotated attributes and relations with a similar scene-graph generation system from (Yang et al., 2020) and (Chen et al., 2020b). In this section, we will describe more details on our data collection and the scene-graph generation method we adopt to generate the referring expressions.

2.1 Data Collection

Among the existed robotic 3D dataset, OCID dataset has sequential object-level scenes that help robots with better knowledge with the instance difference between two subsequent scenes and also the geometric features extracted from point cloud data benefit the robots on vision perceptor (e.g. object grasping or object tracking). Moreover, OCID considers the scenes to be cluttered and we discover that the higher clutter-level will increase the problem difficulty significantly on the grasping task related to the input data format (Appendix A). Hence, according to these investigated advantages, we choose OCID as our original dataset, and extend it with extra semantic annotations such as attributes (e.g., color, texture, shape) and relations (e.g., color relation, spatial relation, etc.) for all the objects in dataset. We design an online web-based annotation tool to collect these extra labels, and we dispatch the labeling tasks over the full-time annotation specialists from a professional data service company. Additionally, we ensure each task is randomly assigned to three trained workers and verified by one checker. The overall tasks take around two months to finish.

2.2 Referring Expression Generation

Gathering the labels we annotated and following the method from the scene-graph based referring
Table 1: Statistic comparison of previous 2D, RGB-D, 3D referring datasets and the OCID-Ref in terms of the number of scenes or images (#Scenes/#Images), number of object categories (#Obj. Cat.), Distractor score (Dis. Score), Data format, number of expressions (#Expressions), and average lengths of the expressions (AvgLen). Our OCID-Ref is the first dataset featuring both 3D signals and object occlusion, which are both crucial for visual grounding for HRI.

|                | # Scenes/# Images | # Obj. Cat. | Dis. Score | Data Format | # Expressions | AvgLen |
|----------------|------------------|-------------|------------|-------------|---------------|--------|
| RefCOCO        | 19,994           | 80          | 4.9        | RGB         | 142,210       | 3.5    |
| Cops-Ref       | 75,299           | 508         | 20.3       | RGB         | 148,712       | 14.4   |
| Sun-Spot       | 1,948            | 38          | 2.46       | RGB-D       | 7,990         | 14.04  |
| ScanRefer      | 703              | 18          | 4.64       | 3D          | 46,173        | 17.91  |
| OCID-Ref       | 2,300            | 58          | 3.36       | 3D          | 267,339       | 8.56   |

Table 2: Show the exemplar templates we use to generate the free-form referring expressions.

| Exemplar Templates                  |
|-------------------------------------|
| Common Sentence                     |
| The <Attr> <Obj>                    |
| The <Attr> object / item.           |
| Relational Sentence                 |
| The <Obj> <Rel>                     |
| The <Attr> <Obj> <Rel>              |
| The <Attr> <Obj1> <Rel> <Obj2>      |

Figure 3: Word cloud of OCID-Ref dataset.

2.3 Dataset Statistics

OCID-Ref uses the same scenes as OCID, containing 2D object segmentation and 3D object bounding boxes for 2300 fully built-up indoor cluttered scenes. Each object is associated with more than 20 relationships with other objects in the same scene, including 3D spatial relations, 2D spatial relations, comparative color relations, and comparative size relations. In Appendix we shows the distribution of different relation mentioned in the referring expression. Moreover, table 1 shows the basic statistic comparison of the previous 2D, RGB-D, 3D referring datasets and the OCID-Ref. To evaluate the difficulty of REC, we follow Cops-Ref to calculate the number of candidate objects of the same categories as the target object (Distractor score) for all scenes. Though there are only 3.36 same candidates in an average of OCID-Ref, lower than 4.64 of ScanRefer, we attribute this difference to the dataset characteristic that our scenes are components of one by one sequence with few objects in the first few scenes. To evaluate the referring performance from no clutter to dense clutter scenes, we follow OCID to separate the scenes into three cluttered levels, free, touching, and stacked, from clearly separated to physically touching to being on top of each other. We also split the val split of ScanRefer into three clutter level.

3 Experiments

We conduct referring expression segmentation experiments on our collected OCID-Ref dataset. We compare different modalities and clutter levels and provide a comprehensive analysis to pave a new path for future research. We also conduct the grasp experiment using different modality data as input.

---

1See Appendix for implementation details
Table 3: Referring expression segmentation performance (pr@0.25) on OCID-Ref. The performance is negatively associated with the clutter level, indicating that the occlusion could make the VG task more challenging. Also, leveraging 3D signals enhances the overall performance on all clutter levels.

Table 4: The performance on ScanRefer in different clutter level.

|        | Free   | Touching | Stacked |
|--------|--------|----------|---------|
| unique | 0.465  | 0.407    | 0.32    |
| multiple | 0.198  | 0.179    | 0.131   |

Table 5: Referring expression segmentation performance on different length of the referring expression.

|        | 2D     | 3D     | Fusion |
|--------|--------|--------|--------|
| short  | 0.508  | 0.592  | 0.645  |
| long   | 0.484  | 0.562  | 0.580  |

3.1 Baseline

We run our experiments with a modern graph-based DGA (Yang et al., 2019) model. We compare 2D (RGB), 3D (Pointcloud) and 2D+3D input signals. For 2D inputs, we use ResNet-101 based Faster-RCNN as our 2D feature extractor and pre-train the extractor on OCID to extract the ROI features from the pool5 layer as the 2D visual features, and use the original DGA's settings for node feature and edge feature on the graph. For 3D inputs, we utilize point-wise features extracted from PointNet (Charles et al., 2017) as the 3D version of the visual feature for each node in the graph. Also, we change the box information from 2D to 3D with box center, box bounds, and box volume. The relations for the edges are modified with 3D relationships between objects instead of 2D relationships.²

To utilize advantages from both 2D and 3D signals, we implement a handy fusion module. We take max-pooling on the point features to aggregate them into a global scene feature and concatenate it to the 2D visual feature as a new visual feature for each object instance, then fuse the box information into (2D box center, 2D box bounds, 2D box area, 3D box center, 3D box bounds, 3D box volume) to preserve the location information from two distinct coordinates. The edge representation is defined as the same as the 3D version.

3.2 Quantitative Analysis

Metric We use Acc@0.25IoU as our metric to measure the thresholded accuracy where the positive predictions have higher intersection over union (IoU) with the ground truths than the thresholds.

Table 3 compares 2D (RGB), 3D (point cloud) models and Fusion model. All models struggle against highly occluded stacked subset while performing well on free subset, indicating that occlusion is a challenge for modern VG models. For

²See Appendix for details
single modality models, the 3D model outperforms and indicates that accurate spatial information is important. Aggregating 2D and 3D signals reach the best performance. Therefore, we suggest future work to explore an effective way to utilize and fuse 2D and 3D signals to tackle our challenging dataset.

3.3 Qualitative Analysis

Figure 4 shows results produced by 2D, 3D baseline, and the fusion model. We discover that all three methods fail when the RE is long and complicated. The fusion method successfully localizes the towel in the scene with 2D and 3D spatial descriptions, while the 3D method has difficulty identifying what is "lower-right". Unsurprisingly, we observe that the 2D method fails on the query with the 3D relation "rear." Figure 4 also shows the failure cases of the fusion method, indicating that our model cannot handle all spatial relations to distinguish between ambiguous objects. 2D and 3D get better performance when the query RE consisted mostly of the common sentences and relationships regarding the whole scene. The failure case suggests that our fusion and localization module can still be improved to utilize the 2D information better and decrease the 3D features’ misuse.

3.4 Ablation Study

Performance in Different Cluttered Level

The clutter distribution of both OCID-Ref and ScanRefer can be seen in figure 1, where OCID-Ref has a relatively balanced quantity of scenes of three different clutter level.

Table 3 shows the result of the model performance in different clutter level. The fused model get the lowest decrease rate consider free scene to stacked scene. We also calculate the the performance in different clutter level for ScanRefer, the result can been seen in the table 4.

Performance on Different Length of Referring Expression

Table 5 shows the result of the model performance on different length of referring expression. We define a short referring expression if its length is smaller than 12 (unit: word of pieces); otherwise, it is a long one. We found the performance drops for all three methods from short to long.

4 Conclusion

In this work, we propose a novel OCID-Ref dataset for VG with both 2D (RGB) and 3D (pointcloud) and occluded objects. OCID-Ref consists of 305,694 referring expressions from 2,300 scenes with providing RGB image and pointcloud inputs. Experimental results demonstrate the difficulty of occlusion and suggest the advantages of leveraging both 2D and 3D signals. We are excited to pave a new path for VG researches and applications.

5 Acknowledgement

This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 110-2634-F-002-026 and Qualcomm Technologies, Inc. We benefit from NVIDIA DGX-1 AI Supercomputer and are grateful to the National Center for High-performance Computing. We also thank Yu-Kai Huang for his insightful suggestion on the figures.

References

Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. 2015. The ycb object and model set: Towards common benchmarks for manipulation research. In 2015 international conference on advanced robotics (ICAR), pages 510–517. IEEE.

R. Charles, Hao Su, Kaichun Mo, and Leonidas Guibas. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. pages 77–85.

Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. 2019. Scanrefer: 3d object localization in rgb-d scans using natural language. arXiv preprint arXiv:1912.08830.

Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K. Wong, and Qi Wu. 2020a. Cops-ref: A new dataset and task on compositional referring expression comprehension.

Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K. Wong, and Qi Wu. 2020b. Cops-ref: A new dataset and task on compositional referring expression comprehension. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. Cognitive Science, 19(2):233–263.

Kuang-Yu Jeng, Yueh-Cheng Liu, Zhe Yu Liu, Jen-Wei Wang, Ya-Liang Chang, Hung-Ting Su, and Winston H. Hsu. 2020. Gdn: A coarse-to-fine (c2f) representation for end-to-end 6-dof grasp detection.
Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. Referit game: Referring to objects in photographs of natural scenes. In *EMNLP*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *CVPR*.

Cynthia Matuszek. Grounded language learning: Where robotics and nlp meet.

Cecilia Mauceri, Martha Palmer, and Christoffer Heckman. 2019. Sun-spot: An rgb-d dataset with spatial referring expressions. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0.

Maria Ralph and Medhat Moussa. 2005. Human-robot interaction for robotic grasping: A pilot study. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 454–459. IEEE.

Mohit Shridhar and David Hsu. 2018. Interactive visual grounding of referring expressions for human-robot interaction. *arXiv preprint arXiv:1806.03831*.

Markus Suchi, Timothy Patten, David Fischinger, and Markus Vincze. 2019. Easylabel: A semi-automatic pixel-wise object annotation tool for creating robotic RGB-D datasets. In (Suchi et al., 2019), pages 6678–6684.

Sibei Yang, Guanbin Li, and Yizhou Yu. 2019. Dynamic graph attention for referring expression comprehension.

Sibei Yang, Guanbin Li, and Yizhou Yu. 2020. Graph-structured referring expressions reasoning in the wild.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. In *Computer Vision – ECCV 2016*, pages 69–85, Cham. Springer International Publishing.
In this Appendix, we provide additional details on the data statistics of the OCID-Ref dataset (Section A); we also provide the detailed setting of our grasping experiment environment in V-REP (Section B), as well as the transformation using the spherical coordinate in figure 6 of relation from 2D to 3D (Section C).

A Qualitative

Additional qualitative result of the referring expression segmentation can be seen in figure 7.

B Grasp Experiment in Simulator

We use V-REP to set up the experimental environment of grasping tasks, which contains the UR5 arm, the RG2 gripper, a table, a box, the objects to grasp, and one depth camera. We show the grasp experiment environment in figure 5.

We produce the point cloud from the depth camera as the input and uses GDN (Jeng et al., 2020) to find the grasps based on the point cloud. We perform a single object grasping experiment in three clutter levels, free, touching, stacked, with 4, 6, 8 objects in the scene, respectively. An object in the YCB (Calli et al., 2015) set is randomly selected and placed on a table, and then the robot tries to grasp the object. If the robot successfully grasps the object from the table to the box, it counts one success. We do 11 trials and calculate the average success rate for each object.

Table 6 demonstrates the results of the single object grasping task in terms of success rate with 2D box and 3D mask in different clutter level. Using a 3D mask rate than a 2D box as an input can get a higher average gripping success rate. The difference increases in a more dense cluttered scene, suggesting that with more accurate segmentation in 3D spatial environment is relatively unaffected in the cluttered and occluded environment.

C Relation transform from 2D to 3D

\[
R(\theta, \phi) = \begin{cases} 
4 & \text{if } \theta < 15 \\
5 & \text{if } \theta > -15 \\
16 + r(\phi) & \text{if } -65 < \theta < -15 \\
8 + r(\phi) & \text{if } 15 < \theta < 65 \\
r(\phi) & \text{otherwise} 
\end{cases} 
\]

(1)

\[
r(\phi) = 6 + \left( \frac{\phi + 22.5}{45} \right) \]

(2)

Table 6: 2D (box) and 3D (mask) Grasp Experiment on different clutter level (%), F: Free, T: Touching, S: Stacked.

|       | top | bottom |
|-------|-----|--------|
| F     | T   | S      | F    | T     | S     |
| 2D    | 65.9| 40.9   | 28.4 | 72.7  | 51.5  |
| 3D    | 72.7| 48.5   | 36.4 | 75.0  | 56.1  |
| \Delta| 6.8 | 7.6    | **8.0** | 2.3   | 4.5   | **10.2** |

Figure 5: Grasp Experiment Environment.

Figure 6: Compute the angles related to 3D relation on spherical coordinates.
Figure 7: Qualitative results from 2D, 3D, and the fusion methods. Predicted masks with an IOU score higher than 0.25 are marked in green, otherwise in red. Examples are tested in the same cluttered scene, and the fusion method produces good results with different REs.