Refer-it-in-RGBD: A Bottom-up Approach for 3D Visual Grounding in RGBD Images Supplementary Material

Haolin Liu[1,2] Anran Lin[1] Xiaoguang Han[1,2,*] Lei Yang[4]
Yizhou Yu[3,4] Shuguang Cui[1,2]
[1]SRIBD, CUHK-Shenzhen[†] [2]FNii, CUHK-Shenzhen[‡]
[3]Deepwise AI Lab [4]The University of Hong Kong

1. Overview

This supplementary material is organized as follows: First, we illustrate the challenges of this proposed task. Second, we describe our newly collected SUNRefer dataset in detail. Third, we furnish more implementation details regarding to our experiment. Finally, we also present more quantitative and qualitative comparison results between our method and the state-of-the-art.

2. Challenges of the task

The challenges of 3D visual grounding in RGBD images are mainly brought by the occlusion problem of RGBD images. As shown in Figure 1, the bed in the RGBD image suffers from severe out-of-view occlusion (self occlusion and mutual occlusion are also common). From this partial observation, it is hard to match this incomplete object with the query sentence. Moreover, it is hard to predict a bounding box that encloses the full shape of the bed since most of the shape information is missing.
Another challenge of RGBD visual grounding is to correctly retrieval the object during on-the-fly visual grounding on RGBD streaming data. Since it also requires distinguish the target object with all other objects presented in the environment, which have equivalent difficulties as 3D visual grounding in a complete scene.

3. Dataset

SUNRefer dataset has 38,495 annotations of 7,699 objects. Each object is given 5 different descriptions concerning its category, appearance and spatial relationship with neighboring objects.

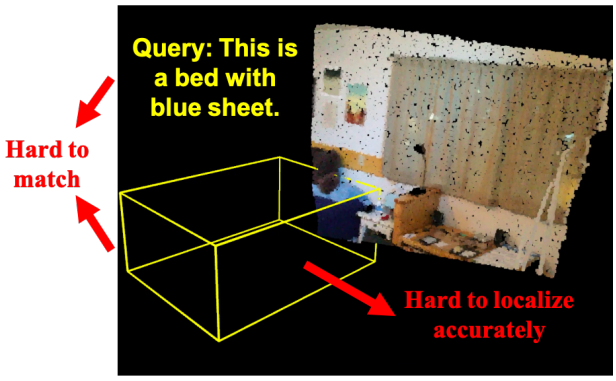


Figure 1. Illustration of the occlusion problem of RGBD images. The yellow bounding box is the target ground truth.

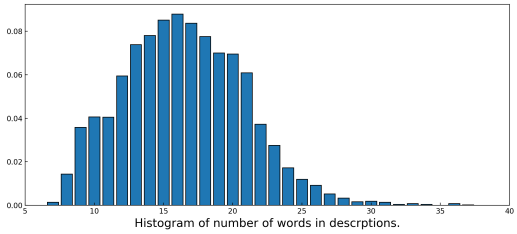
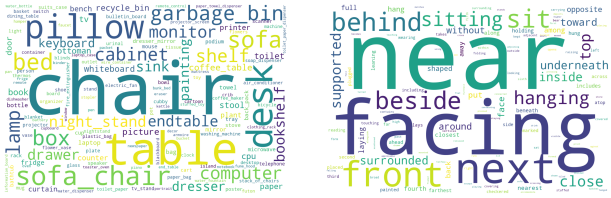


Figure 2. Distribution of number of words in SUNRefer dataset.



(a) (b)

Figure 3. Word cloud of (a) object names (b) spatial relations

*Corresponding Email: hanxiaoguang@cuhk.edu.cn
†Shenzhen Research Institute of Big Data
‡The Future Network of Intelligence Institute, The Chinese University of Hong Kong, Shenzhen

3.1. Dataset collection

We hire professional annotators to annotate each target object with referring expressions. For each target object, an RGB-D image with a bounding box enclosing this object is given and the annotators should describe the object from various aspects. We ask annotators to pay attention to object's appearance such as color, shape, size and material. Detailed descriptions are recommended. We also ask the annotator to additionally describe the spatial relationship between the target objects and their neighboring objects. Each description should be able to uniquely identify the target object. To further ensure diversity of the annotations, we assign different annotators to describe the same object.

Verification is performed after the annotation. We hired different workers than the annotators to check the annotation contents. The descriptions that fail to identify the target object are filtered. We further correct spelling and grammar issues of the descriptions.

3.2. Dataset Statistics

SUNRefer dataset is based on 7,699 images from [6]. To include more detailed information as well as making the language more diverse, there is no word count limit and the distribution of number of words is shown in Figure 2. Word cloud of object names in the SUNRefer dataset is shown in 3(a). Besides, Figure 3(b) shows the word cloud of spatial relations in the SUNRefer dataset. Some examples from the SUNRefer dataset are shown in Figure 4.

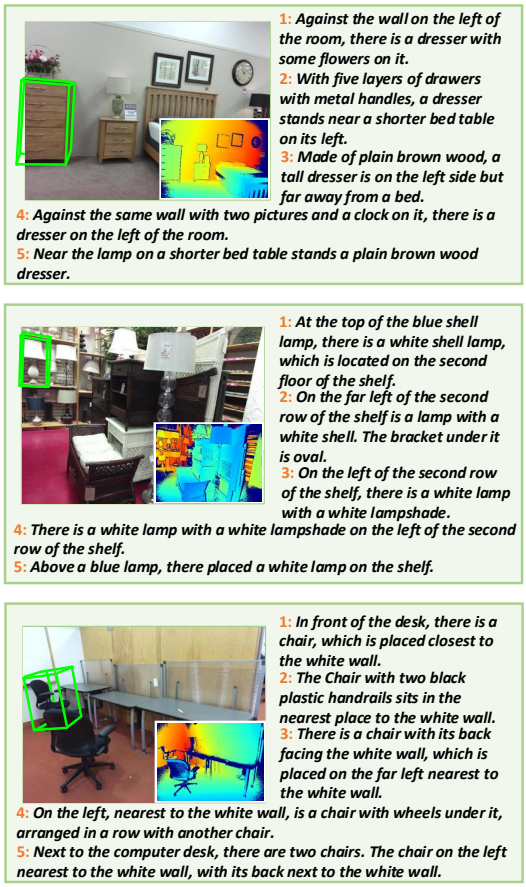


Figure 4. Some examples from SUNRefer dataset.

4. More implementation details

In this section, we present some detailed information of the network architecture of some critical components, full algorithm of the weighted FPS, and experiment detail of the image and point cloud modality in ablation study.

4.1. Network architecture

The heatmap generation model takes as input the 256-dim language feature l and a voxelized RGBD image with 3 input channels (only RGB). The model is a 3D U-Net structure with additional visio-linguistic fusion at the bottleneck layer. The convolution and transposed convolution are implemented as sparse convolution [2]. The network architecture is shown in Figure 5. The parameters of the convolution layer are the size of cubic kernel, the number of input channels, and the number of output channels, respectively. At the bottleneck layer, the language features l is first concatenated with the feature map, then visio-linguistic fusion is conducted.

In Adaptive Feature Learning, Pointnet++ encoder takes as input a point cloud with 20,000 points to generate a set of 512 keypoints with features. We use three Set Abstraction (SA) layers [4] to downsample and encode the point clouds. The parameters for each layer is shown in Table 1. 256 seed points are sampled via the weighted FPS from the point cloud, followed by feature aggregation from these 512 keypoints.

Seed point features are fused with language feature l using another visio-linguistic fusion module. The seed point features and the language feature l are firstly concatenated forming a 512-dim vector and input to another visio-linguistic fusion module. Then, it is followed by three layers of convolutions with ReLU activation using kernel size $= 1$. Both the numbers of input and output channels are fixed, i.e. 512. The voting module has exact the same architecture with [3] except that the input feature dimension is 512. After the voting module, the framework finally yields 32 proposals and matching scores.

4.2. weighted FPS

Weighted FPS replaces the Euclidean distance metric in FPS with a weighted generalization of Euclidean distance

layer	SA1	SA2	SA3
# of sub-samples	2048	1024	512
grouping radius	0.1	0.2	0.4
# of grouping samples	64	32	16
MLP processing	$3 \times 64 \times 64 \times 128$	$128 \times 128 \times 128 \times 256$	$256 \times 256 \times 256 \times 256$

Table 1. Parameters of Set Abstraction layer in Pointnet++ encoder.

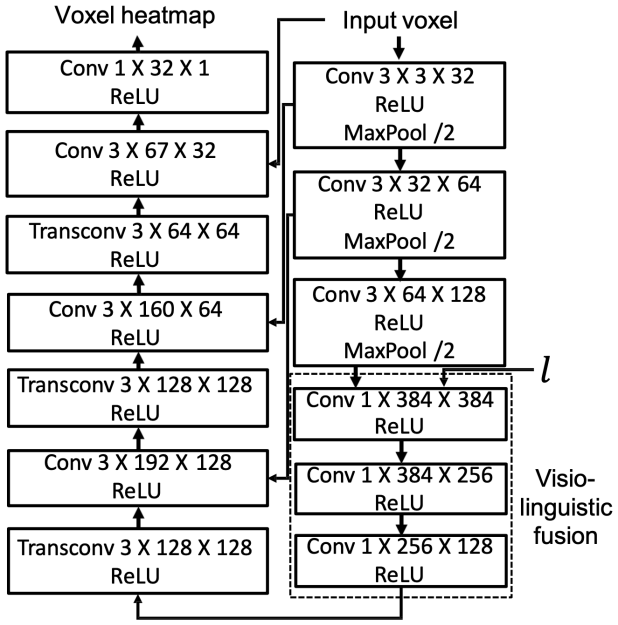


Figure 5. Network architecture of the voxel heatmap generation model.

Algorithm 1 Weighted Farthest Point Sampling

Input: Point cloud $\mathbf{P} \in \mathbb{R}^{N\times 3}$, heatmap $\mathbf{H} \in \mathbb{R}^{N\times 1}$
Output: Seed points $\mathbf{s} \in \mathbb{R}^{M\times 3}$

$\mathbf{s}_0 = \mathbf{P}_0$
$i = 1$
while $i < M$ **do**
  $\mathbf{c} = \mathbf{s}_{i-1}$
  $j' = \arg\max_j \mathbf{H}_j d(\mathbf{P}_j, \mathbf{c})$
  $\mathbf{s}_i = \mathbf{P}_{j'}$
  $i = i + 1$
return $\mathbf{s}$

metric as stated in the paper:

$$\hat{d}(\mathbf{q}, \mathbf{c}) = h(\mathbf{q})d(\mathbf{q}, \mathbf{c}). \quad (1)$$

We use the iterative algorithm in FPS to sample seed points from the point cloud. The whole algorithm is stated in Algorithm 1. The implementation is parallelized using CUDA for efficient computation.

4.3. Image and point cloud modality

In the ablation study, we discussed three heatmap generation models using different modalities, i.e., voxel, image, and point cloud. We provide more details about the image and point cloud modalities here. For the image modality, U-Net [5] is used to process 2D images to obtain an image-based heatmap which is later mapped to the point cloud using a similar procedure as is done in our framework. The variant using point cloud also employs a U-Net like architecture similar with [4] and generates a point cloud heatmap directly. For both modalities, visio-linguistic fusion is conducted at the bottleneck layer.

5. More experiment

5.1. Different Occlusion Ratio

We also examine how the occlusion will affect the performance, and compare with previous method [1]. We conduct this experiment on ScanRefer(single-RGBD) dataset. Quantitative experiment is shown in Table 2. We defines the occluded ratio as the ratio between the non-overlapped volume of the partial and the full bounding box, and the volume of the full bounding box. It is reasonable that with higher occlusion rate, the retrieval accuracy drops significantly. However, our method significantly outperforms [1] no matter what the occluded ratio is.

Metric	Occluded ratio	<20%	20-40%	40-60%	>60%
Acc@0.5	Ours	47.0	36.6	24.9	12.7
Acc@0.5	ScanRefer[1]	29.9	22.7	15.8	8.69
Acc@0.25	Ours	67.1	60.4	55.5	45.7
Acc@0.25	ScanRefer[1]	56.9	50.2	43.7	35.6

Table 2. Quantitative comparison under different occluded ration.

5.2. 'Unique' and 'Multiple' subset

We also conduct experiment on 'unique' and 'multiple' subset as in [1], where image in 'unique' subset only contains one unique object that have the same class as the target object; image in 'multiple' subset contains multiple objects having the same class. The quantitative results is shown in Table 3. Our method significantly outperforms [1] on both subsets.

Methods	unique			multiple		
	Acc@0.5	Acc@0.25	R@5	Acc@0.5	Acc@0.25	R@5
ScanRefer[1]	21.1	51.7	33.1	18.9	35.2	34.1
Ours	33.8	64.2	48.9	24.1	36.9	51.1

Table 3. Quantitative comparison on different subset. The experiment is conducted on ScanRefer (single-RGBD).

6. More results

In the paper, we compare our method with ScanRefer [1], One Stage [8], and ReSC [7]. [8] and [7] are both one-stage 2D visual grounding methods that generate both proposals and matching scores. In order to lift the 2D grounding results to 3D space, we additionally train a VoteNet detector [3] to generate 3D proposals. Then, we project these 3D proposal to the 2D image plane. The final 3D grounding results are considered as those having the largest overlapping regions with the 2D grounding results. More qualitative comparison results are shown in Figure 6 and Figure 7. All example in these two figures are from the test set of our SUNRefer dataset. Ours methods have much higher retrieval accuracy than the other methods since they often fail to retrieve the correct objects. Even when the other methods successfully retrieve the correct object, our method still outperform them in terms of localization accuracy. For example, in the second row in both Figure 6, Figure 7 or the last row in Figure 7, ScanRefer retrieve the trash bin with a losen bounding box while our method produce a more tight bounding box.

References

[1] 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.

[2] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019.

[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 International Conference on Computer Vision*, pages 9277–9286, 2019.

[4] 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*, pages 652–660, 2017.

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[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*, pages 567–576, 2015.

[7] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *ECCV*, 2020.

[8] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *ICCV*, 2019.

Figure 6. Comparison between our method and the state-of-the-art. A bounding box is considered as a successful prediction (green) if it has an IoU larger than 0.5 with the ground-truth box (yellow); otherwise, it is considered as a failure one (red).

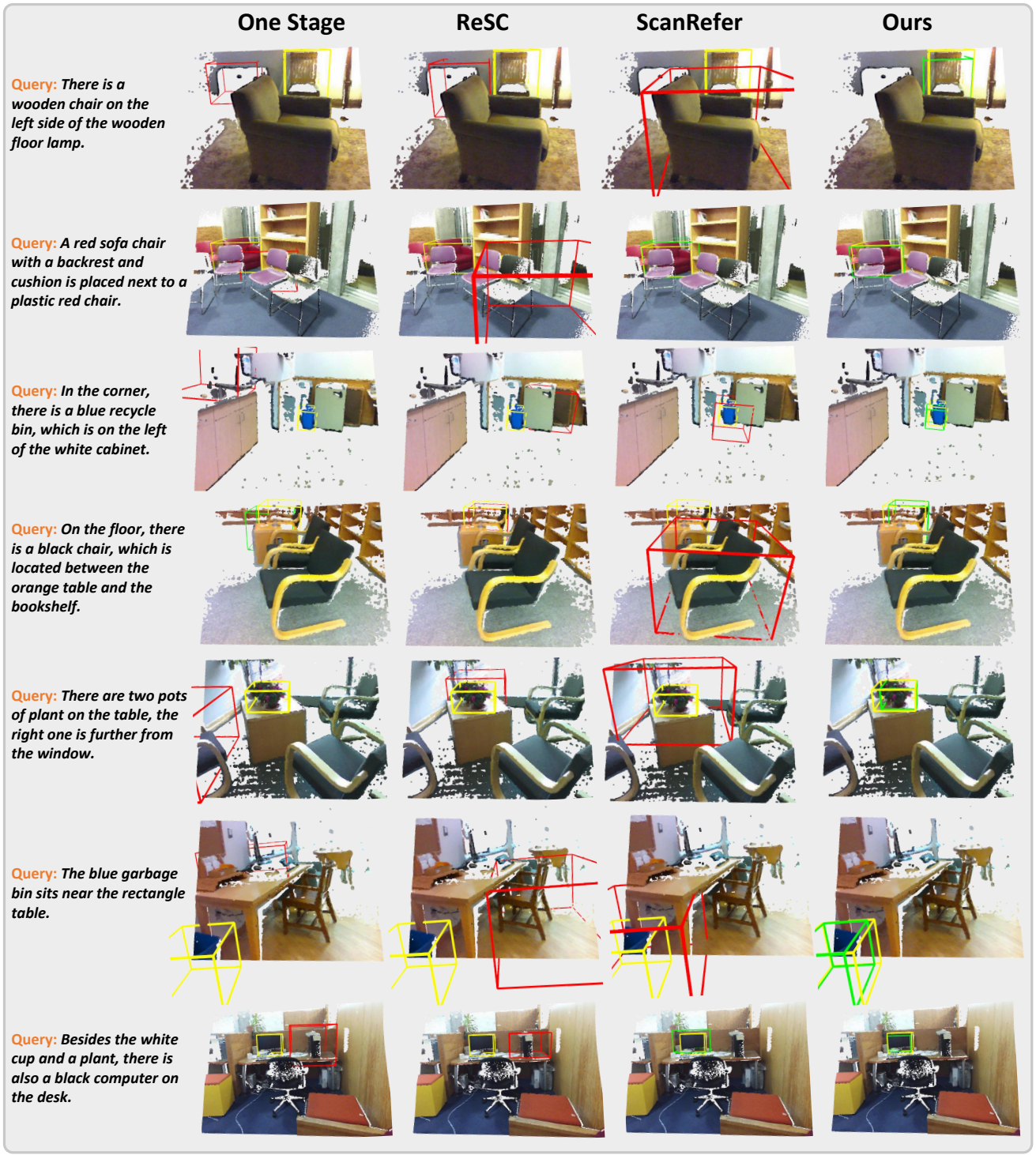


Figure 7. Comparison between our method and the state-of-the-art. A bounding box is considered as a successful prediction (green) if it has an IoU larger than 0.5 with the ground-truth box (yellow); otherwise, it is considered as a failure one (red).