Free-form Description Guided 3D Visual Graph Network for Object Grounding in Point Cloud (Supplementary Material)

We provide more model detail and experimental results in this supplementary material. In detail, we provide more loss function detail in Section A, we provide more ablation studies in Section B. Section C shows more qualitative results on ScanRefer [2] and Nr3D [1] dataset. Section D shows the enlarged figure of model visualization.

A. Loss function

We compute a vector representation for relation phrase as the edge features between node v_i^l and v_j^l in language scene graph $\mathcal{G}^l$. The edge features in 3D visual graph $\mathcal{G}^u$ are extracted from the 3D points within the minimum box region, which denote the relation features between the node v_i^u and v_j^u, and correspond to the edge features in language scene graph $\mathcal{G}^l$.

The final loss of our model is a linear combination of the vote loss $\mathcal{L}_{vt}$ [3], abjectness of loss $\mathcal{L}_{obj}$, bounding box loss $\mathcal{L}_b$, semantic classification loss $\mathcal{L}_{sm}$, description classification loss $\mathcal{L}_{cls}$ and reference loss $\mathcal{L}_{rf}$. We provide the implementation detail for the reference loss $\mathcal{L}_{rf}$ in this supplementary material,

$$\mathcal{L}_{rf} = \mathcal{L}_{rf}^1 + \mathcal{L}_{rf}^2 + \mathcal{L}_{rf}^3, \tag{1}$$

where $\mathcal{L}_{rf}^1$ supervises the matching score estimation process in the nodes pruning step, $\mathcal{L}_{rf}^2$ supervises the bounding box offset regression process in the selected proposals refinement step, and $\mathcal{L}_{rf}^3$ the matching score estimation process in the prediction model.

In the nodes pruning step, the labels for matching score are defined as soft distributions based on the IoU between the K_o 3D bounding box candidates and the ground truth 3D bounding box of target object. $\mathcal{L}_{rf}^1$ can be described as:

$$\mathcal{L}_{rf}^1 = -\sum_{k=1}^{K_o} \left[\ln\left[Softmax(P_s^1)\right] \cdot l_1\right], \tag{2}$$

where P_s^1 denotes the predicted matching scores, l_1 presents the ground truth labels, K_o is the number of 3D bounding box candidates.

In the refinement step, the labels for offset regression are computed using the selected 3D bounding boxes and the ground truth 3D bounding box of target object. $\mathcal{L}_{rf}^2$ can be described as:

$$\mathcal{L}_{rf}^2 = \sigma L_{sm}(R_r, R_{gt}), \tag{3}$$

where R_r and R_{gt} present the predicted offset and its ground truth respectively, σ is a coefficient and set to 0.1, L_{sm} denotes the Smooth-L1 function.

Similarly, the labels are defined as soft distributions based on the IoU between the K refined 3D bounding box candidates and the ground truth 3D bounding box of target object in the prediction model. $\mathcal{L}_{rf}^3$ can be described as:

$$\mathcal{L}_{rf}^3 = -\sum_{k=1}^{K} \left[\ln\left[Softmax(P_s^2)\right] \cdot l_2\right], \tag{4}$$

where P_s^2 denotes the predicted matching scores, l_2 presents the ground truth labels, K is the number of selected 3D bounding box candidates and set to 20.

B. Extra ablation studies

For ScanRefer, we measure the percentage of predictions whose IoU with the ground truth is greater than 0.25 and 0.5. If there is only a single object of its class in the scene, we take it as unique, otherwise multiple. For Nr3D, the models are evaluated by accuracy i.e., whether the model correctly selects the referred object from the M proposals.

In Table 1, we also perform experiments to show the impact of the selected proposals number K in 3D visual graph model. We can see that the accuracy of 3D object grounding improves steadily using more relation graphs (while keeping everything else constant) up to $K = 20$ after which there is a slight drop in accuracy. Therefore, we use $K = 20$ in the remaining experiments. An intuitive explanation is that, when K is too small, the 3D proposals related to the target will be missed while matching the language scene graph nodes with the noisy 3D bounding box candidates. However, if we continue to increase the value of K, our model will get a comparable performance but it will cause too many nodes in the subsequent 3D visual graph model, and the number of model parameters also increases leading to over-fitting.

C. Extra qualitative results

Figure 1 shows extra 4 qualitative visual grounding results produced by the ScanRefer [2] method and our method on the ScanRefer dataset [2] (1-3 columns) and the Nr3D [1] dataset (last column). From the final outputs, it can be observed that our proposed method generates better results than ScanRefer [2].

Methods	Unique Acc@0.5	Multiple Acc@0.5	Overall Acc@0.5
K=5	59.82	16.49	25.04
K=10	64.35	22.91	30.24
K=15	67.10	25.20	33.66
K=20	**67.94**	**25.70**	**34.01**
K=25	67.87	25.64	33.82

Table 1. Comparing the affect of the selected proposals number K.

	VoteNet	ScanRe.	InstanceRe.	Ours
Param.	11.2 M	14.4 M	15.7 M	14.6 M

Table 2. Model complexity comparison with SOTA methods.

D. Model visualization

We show the enlarged Figure 4 in manuscript, as shown in Figure 2.

E. Model Complexity

The complexity of our network (Table 2) is much lower than InstanceRefer [4] and only slightly more than ScanRefer [2] and VoteNet [3].

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European Conference on Computer Vision*, 2020. 1, 3

[2] 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*, 2020. 1, 2, 3

[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*, pages 9277–9286, 2019. 1, 2

[4] Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Zhen Li, and Shuguang Cui. Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. *arXiv preprint arXiv:2103.01128*, 2021. 2

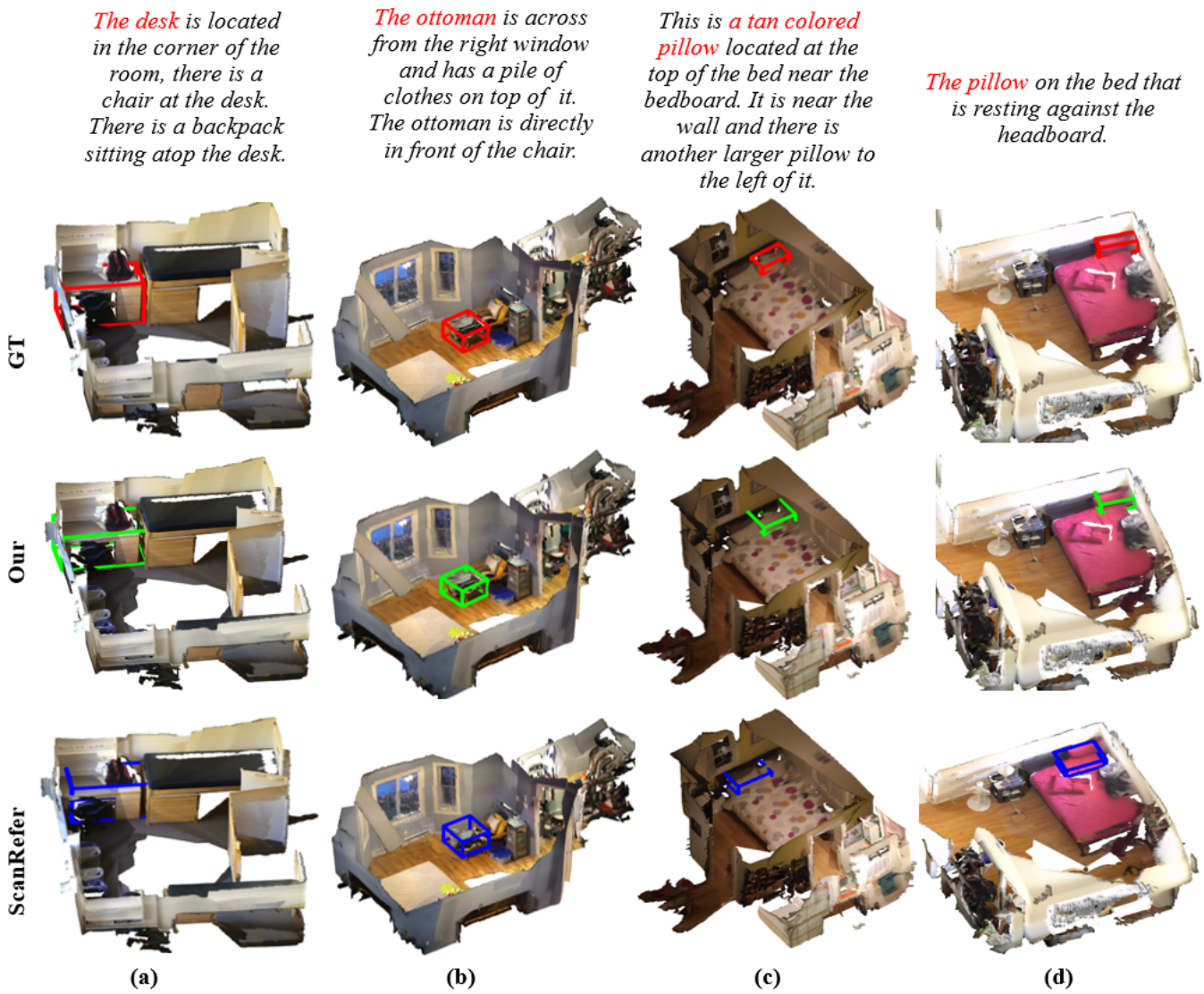

Figure 1. Results of ScanRefer [2] method and our method on ScanRefer [2] dataset (columns 1-3) and Nr3D [1] dataset (last column).

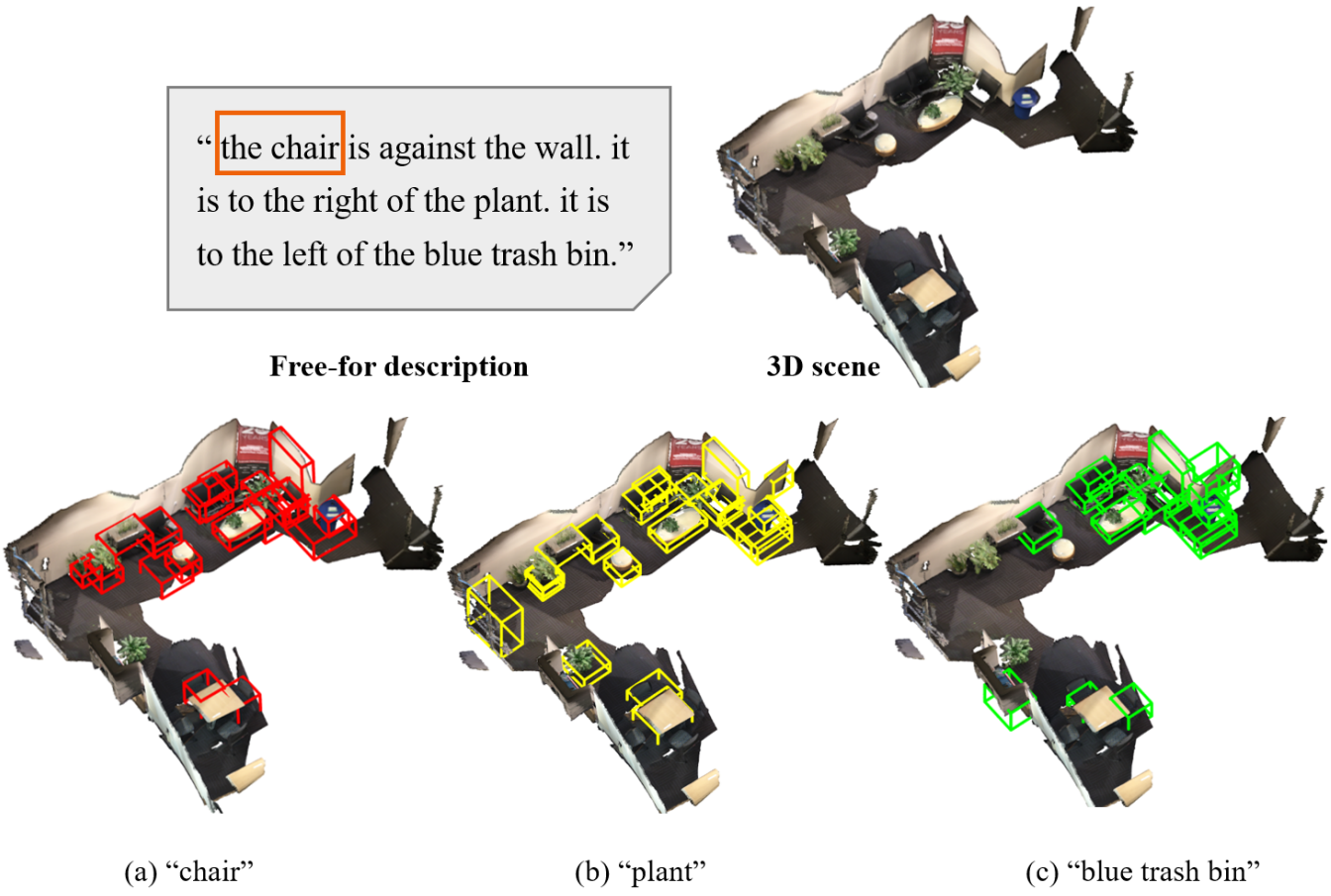

Figure 2. Results of the most relative 3D bounding boxes foreach noun phrase in description guided 3D visual graph module.