

Supplementary Materials: Cross-Task Knowledge Transfer for Semi-supervised Joint 3D Grounding and Captioning

Anonymous Authors

1 EXPERIMENTS ON MORE DATASETS

To further validate the effectiveness of our proposed 3D-CTTSF method, we conducted additional experiments on the Nr3D [1] dataset. In our experiments, we adopted a similar experimental setup to ScanRefer, wherein we localize objects relevant to the given text in the scene without providing ground truth bounding boxes for the objects. We utilized Acc@0.25 as the metric. Comparative analysis was conducted against some methods that reported results under the same experimental settings, as shown in Table 1. To fairly compare the existing methods in the same semi-supervised setting, we also re-implement some approaches by using only 10% labeled data. The results indicate that our approach outperforms most fully supervised methods using 100% annotated data, though slightly inferior to some state-of-the-art methods. Furthermore, when using an equal amount of annotated data, our approach significantly outperforms all previous methods.

We did not utilize another dataset, Sr3D, provided by ReferIt3D [1], as the text in Sr3D is synthetically generated based on rules rather than manually annotated. While such a dataset may be suitable for 3D visual grounding, it is not appropriate for 3D dense captioning tasks. Since our approach jointly trains these two tasks, we chose not to employ the Sr3D dataset.

Table 1: Results of 3D visual grounding on Nr3D dataset.

Method	label	Acc@0.25
ReferIt3D [1]	100%	24.0
LanguageRefer [2]	100%	28.6
InstanceRefer [4]	100%	29.9
SAT [3]	100%	31.7
ReferIt3D [1]	10%	11.2
3DVG-Transfomer [5]	10%	12.8
InstanceRefer [4]	10%	21.4
SAT [3]	10%	19.7
Ours	10%	30.6

2 ABLATION STUDY ON ASYMMETRIC DATA AUGMENTATION

We conducted an ablation study on the asymmetric data augmentation employed in our method, with results summarized in Table 2. The first two columns of the table denote the types of data augmentation used for the input of the teacher model and the student model, respectively. 'Weak' indicates the use of only resampling, while 'strong' indicates a combination of resampling, random flipping, random translations, random size scaling, and random rotation. The outcomes presented in the table demonstrate that asymmetric data augmentation enhances the performance of pseudo-label learning. Employing the same level of augmentation (either strong or weak)

for both teacher and student models undermines training effectiveness. This is because such a practice does not guarantee that the teacher's predictions surpass those of the student, thus reducing the effectiveness of student learning from the teacher.

Table 2: Results of 3D visual grounding on Nr3D dataset.

Augmentation		Acc@0.25	Acc@0.5
teacher	student		
weak	weak	45.91	33.08
weak	strong	47.97	35.28
strong	strong	47.37	34.21

3 MORE VISUALIZATION RESULTS

To comprehensively demonstrate our experimental results, we conducted additional visualization experiments using 10% of annotated data, as shown in Figure 1.

REFERENCES

- [1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. 2020. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 422–440.
- [2] Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. 2022. LanguageRefer: Spatial-language model for 3d visual grounding. In *Conference on Robot Learning*. PMLR, 1046–1056.
- [3] Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. 2021. Sat: 2d semantics assisted training for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1856–1866.
- [4] Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Sheng Wang, Zhen Li, and Shuguang Cui. 2021. Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1791–1800.
- [5] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 2021. 3DVG-Transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2928–2937.

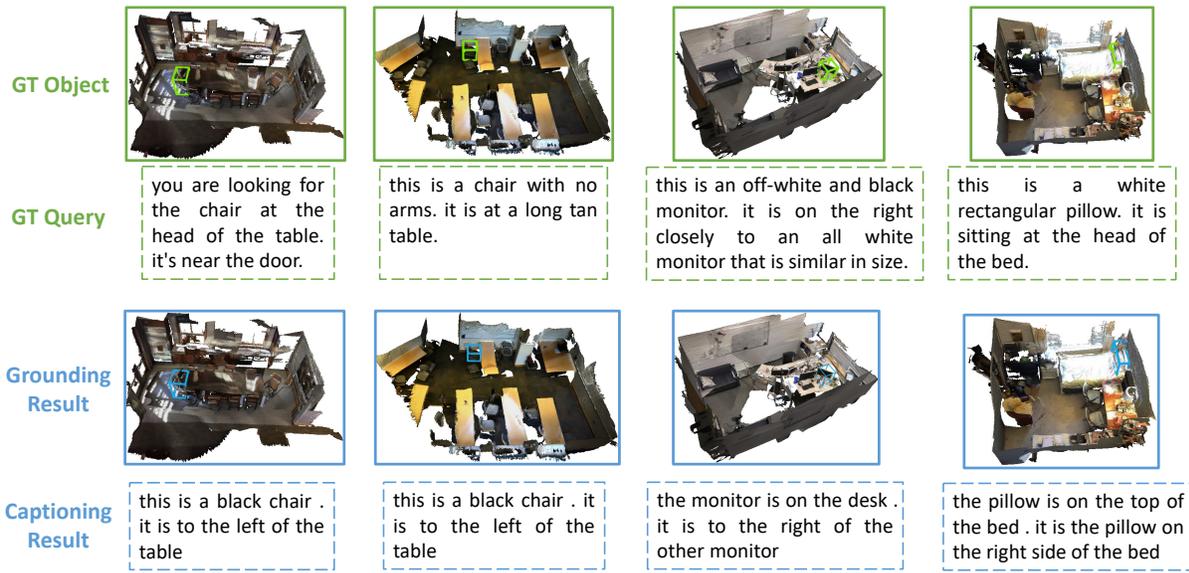


Figure 1: Visualization of more results. The first two rows represent the ground truth of object positions and the query text, respectively. The third row presents the results of 3D visual grounding, while the last row presents the results of 3D captioning.

Unpublished work
Not for distribution