Supplementary Materials for Fine-Grained Spatial and Verbal Losses for 3D Visual Grounding

Sombit Dey[1,2] Ozan Unal[1,3*] Christos Sakaridis[1] Luc Van Gool[1,2,4]
[1]ETH Zurich, [2]INSAIT, [3]Huawei Technologies, [3]KU Leuven

Method	Overall
Top-down	58.9
Bottom-up	56.3

Table 1. Comparison of top-down and bottom-up attention masking for the verbo-visual fusion module

1. Dataset

We evaluate our method on the Nr3D and Sr3D datasets [2] that provide human-annotated and synthetically generated utterances to identify instances on ScanNet [4]. Specifically, Nr3D consists of 41,503 natural language queries and Sr3D consists of 83,572 descriptions over 707 unique indoor scenes referring to 76 object classes. For evaluation, the dataset is split in two ways. First split considers the difficulty of the grounding task, with referred objects that have only 2 distractors considered *easy*, and objects that have up to 6 distractors are considered *hard*. Second form of splitting considered view-dependency. Based on distinct word mining, utterances are classified as either view-dependent or view-independent.

2. Results and Evaluations

2.1. Multiple Objects

Description referring to multiple objects is a compelling research problem that the ReferIt3D benchmark doesn't address. We carry out further analysis of our approach using the ScanEnts dataset [1]. to evaluate the network performance with the number of anchors in the utterance 1. A drop in performance with more objects in the utterance is observed.

3. Top-Down Masking

To explicitly limit information routing within local neighborhoods, ConcreteNet [6] employs a masking operation that hinders self-attention between spatially distant instance candidates. This masking operation is implemented in a bottom-up manner where the masking sphere grows as the layers increase. To be able to utilize the offset loss introduced in the main manuscript at various stages of the bidirectional attentive fusion module, we instead use a top-down approach, allowing all instances to be aware of one another even at the shallowest levels. In Tab. 1, we ablate this masking approach and show that a top-down method works better than a bottom-up one.

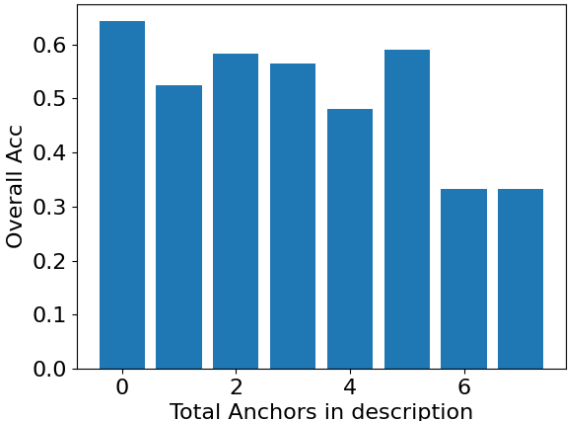

Figure 1. Plot visualizing the overall accuracy against the length of the natural language description #tokens.

3.1. Out of Distribution

We address the out-of-distribution ability of AsphaltNet. In table 2 we evaluate our network on the scanrefer dataset [3] with GT boxes provided in the input with network weights corresponding to NR3D dataset. The network trained without any of the contribution of AsphaltNet is used as baseline for the comparisons. We observe decent performance on the scanrefer dataset given the vast difference in the language domain of ReferIt3D and scanrefer datasets showcasing the OOD robustness of AsphaltNet. We further investigate the synthetic to real transferability of the network. AsphaltNet trained on SR3D dataset is tested on the NR3D dataset to test for synthetic - real do-

*Corresponding author: Ozan Unal, ozan.unal@vision.ee.ethz.ch

Method	Accuracy
Baseline	43.89
AsphaltNet	47.99

Table 2. Evaluation of OOD performance on ScanRefer

Method	Accuracy
Baseline	34.87
AsphaltNet	38.53

Table 3. Evaluation of OOD performance on NR3D

main adaptability. AsphanltNet yields relatively lower performance , table 3, when evaluated on NR3D dataset. This can attributed to a lack of diversity in the synthetic dataset hence unable to transfer to the natural language settings.

3.2. Multi-View Ensembling

MVT [5] proposes a multi-view feature aggregation approach that aggregates visual features from different viewpoints during both training and testing. While this acts as a form of regularization during training with the goal of smoothing the gradients, during test time, the multi-view aggregations act as a form of test-time-augmentation (TTA). AsphaltNet's performance can be enhanced by utilizing TTA by employing multiple forward passes through affine-transformed point clouds. Specifically, we employ multi-view ensembling (MVE) following ConcreteNet [6] where we infer a scene N times, each with the point cloud rotated by an angle $r \in [0, 2\pi]$. The final prediction is determined through majority voting across the N inferences (with $N = 9$). As seen in table 4, MVE further boosts the performance of AsphaltNet for both NR3D and SR3D splits of ReferIt3D.

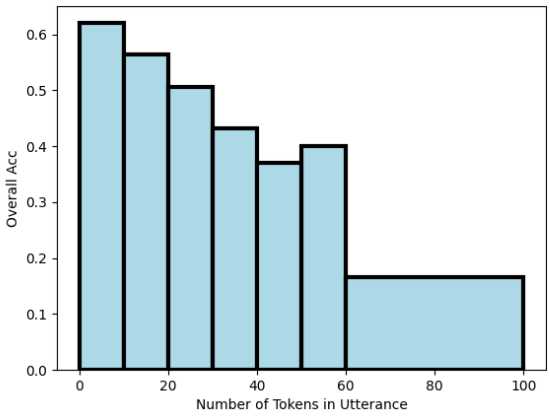

Figure 2. Plot visualizing the overall accuracy against the length of the natural language description #tokens.

4. Limitation and Discussion

While AsphaltNet achieves state-of-the-art results on both ReferIt3D [2] benchmarks, we still observe an important limitation when analyzing the predictions. In Fig. 2 we show the accuracy of our model on the Nr3D *val*-set against the number of tokens in the natural language prompt. As seen, we observe a linear drop in performance as the number of tokens increases. We speculate this is due to the increased difficulty of the span prediction task that limits our method's generalization ability for longer text descriptions.

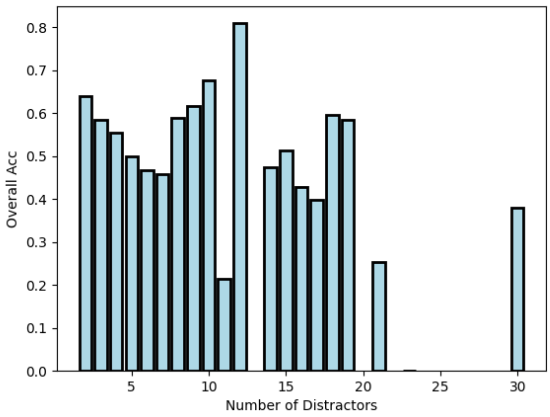

Figure 3. Plot visualizing the overall accuracy against the total number of anchors in the description.

Furthermore, in Fig. 3, we show that while our method's performance weakens as the number of tokens increases (verbal input), the same behavior is not observed when investigating the visual input changes, i.e. as the number of same-class distractors increases. This illustrates the benefits of the offset loss, where the increased robustness towards spatial localization reduces the negative effects of increased same-class instances within the scene.

References

[1] Ahmed et al. Abdelreheem. Scanents3d: Exploiting phrase-to-3d-object correspondences for improved visio-linguistic models in 3d scenes. 1

[2] 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. 1, 2

[3] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[4] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 1

	Network	Overall	Easy	Hard	View-Dep	View-Ind
NR3D	AsphaltNet	58.9	64.3	53.9	57.4	59.8
	AsphaltNet + MVE	**59.3**	64.9	54.2	57.8	59.9
SR3D	AsphaltNet	69.7	71.9	64.5	67.2	70.0
	AsphaltNet + MVE	**72.7**	74.6	68.2	70.4	73.0

Table 4. Evaluation of AsphaltNet with test-time-augmentation (MVE)

[5] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3D visual grounding. In *CVPR*, 2022. 2

[6] Ozan Unal, Christos Sakaridis, Suman Saha, and Luc Van Gool. Four ways to improve verbo-visual fusion for dense 3d visual grounding. In *European Conference on Computer Vision (ECCV)*, 2024. 1, 2