SLiP: Situated-oriented Localization in Point Clouds

Shaocong Xu[1,2], Xiaoxue Chen[2,3], Hao Zhao[3], Guyue Zhou[3]
[1]School of Informatics, Xiamen University
[2]Institute for AI Industry Research (AIR), Tsinghua University
[3]Department of Computer Science and Technology, Tsinghua University
xushaocong@stu.xmu.edu.cn, {chenxiaoxue,zhaohao,zhouguyue}@air.tsinghua.edu.cn

1. Introduction

Situation Understanding involves localizing the position and quaternion referred to in a situation description within a given scene. This task is essential for agents to make informed decisions based on the context and environment, such as for situated reasoning [4].

Despite the importance of this task, there is currently no specific method for achieving it. This gap in the literature highlights the need for novel approaches that can effectively address this challenge. To this end, we propose SLiP for Situation Understanding that leverages non-parametric situation-oriented queries to accurately localize the position and quaternion within a given scene in a gradual manner.

We evaluate our method on SQA3D [4] and demonstrate its effectiveness in achieving high accuracy and robustness in Situation Understanding.

In summary, we propose a new method, called as SLiP, for Situation Understanding and achieve promising results on SQA3D [4]

2. Method Overview

Inspired by the promising 3D language grounding results achieved by BUTD-DETR [1] and the similarity between the 3D language grounding and Situation Understanding tasks, which both involve localizing an object/situation referred to by an utterance within a given scene, we selecte BUTD-DETR as our baseline and adapted it for the task of Situation Understanding.

The task of Situation Understanding involves localizing the position and quaternion referred to in a situation description D within a given scene represented as a point cloud $P \in \mathbb{R}^{N\times 6}$.

Feature Extraction: As shown in fig. 1, the P and D are passed through PointNet++ [5] and RoBERTa [2] seperately to extracting feature, attaining point cloud feature $P_{\text{seed}} \in \mathbb{R}^{M\times(3+C)}$ and text feature $L_{\text{seed}} \in \mathbb{R}^{W\times C}$.

Multi-Modal Cross Encoder: Afterward, a multi-modal cross encoder with N_E layers, each layer composed of self- and cross-attention sequentially, extracts cross-modal features between P_{seed} and L_{seed}, resulting in the generation of text-aware point cloud feature $P_1 \in \mathbb{R}^{M\times(3+C)}$ and scene-aware text feature $L_1 \in \mathbb{R}^{W\times C}$.

Situation-oriented Decoder: In order to filter the text-aware point cloud out and generate a situation-orientet queries for decoder, we employ a MLP to compute the probability of the point referred to by the description, resulting in score S. The point clouds with top k highest score are selected, which can be expressed as:

$$\begin{aligned} S &= \text{MLP}(P_1), \\ P_{\text{query}} &= P_1[\text{argtopk}(S,k)], P_{\text{query}} \in R^{k\times(3+C)}. \end{aligned} \tag{1}$$

Finally, a situation-oriented decoder with N_D layers, each consisting of a self-attention, two cross-attention, and a prediction head sequentially, is used to mine the situation cue between P_1 and L_1, and predicts position $\hat{p} \in R^3$, quaternion $\hat{q} \in R^4$ and corresponding query score $\hat{s}_q$.

Loss Function: As each layer of the decoder is crucial for situation understanding, we supervise the output of all situation-oriented layers. For each decoder, we filter the best prediction by $\hat{s}_q$ and supervise it with an Euclidean norm loss and MSE loss. As the prediction of the z value in $\hat{p}$ is irrelevant for this task, we do not supervise it. The formulation can be writed as:

$$\begin{aligned} \mathcal{L}_{\text{qua}} &= \sqrt{\sum_{i=1}^{4} |\hat{q}_i - q_i|^2}, \\ \mathcal{L}_{\text{pos}} &= \sqrt{(\hat{p}_x - p_x)^2 + (\hat{p}_y - p_y)^2}, \\ \mathcal{L}_{\text{total}} &= \lambda_{qua}\mathcal{L}_{\text{qua}} + \lambda_{pos}\mathcal{L}_{\text{pos}}. \end{aligned} \tag{2}$$

The p and q represent the ground truth. λ_{qua} and λ_{pos} are hyperparameters used for the quaternion loss and the position loss, respectively.

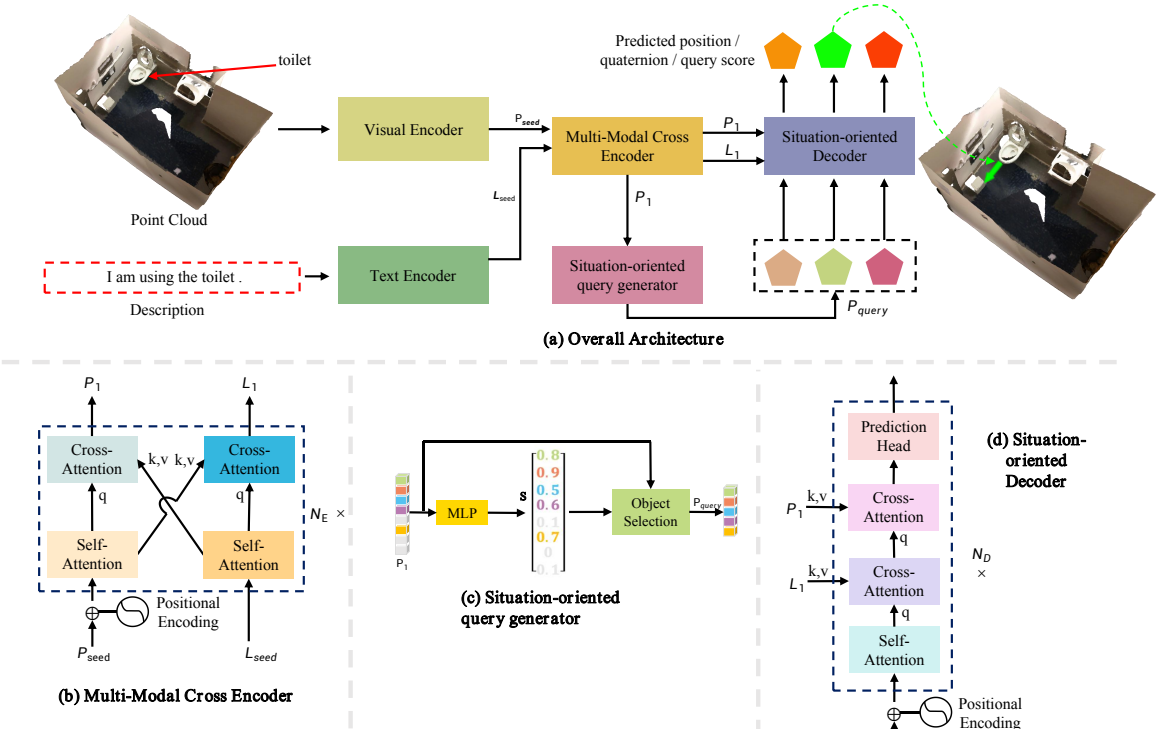

Figure 1. Pipeline of our method.

3. Experiments

3.1. Implementation Details

we utilized one GeForce RTX 3090 to train our model for 200 epochs with a batch size of 16. We employed the AdamW optimizer [3] with an initial learning rate of $1e-4$ to improve the convergence of the optimization process. We set N_E and N_D to 3 and 6, respectively. The value of k was set to 256. both of λ_{qua} and λ_{pos} are set to $1e+1$.

3.2. Comparison with state-of-the-art

In this study, we evaluated the effectiveness of our algorithm on the SQA3D dataset [4]. Our experimental results, as presented in table 1, clearly demonstrate the superiority of our method over the previous state-of-the-art solution by a large margin.

However, we also observed that the increase in the $Acc@30°$ metric was not significant, which we attribute to the ease of the evaluation metric. Specifically, the evaluation metric considers a prediction to be correct if the difference in the z-axis rotation angle between $\hat{q}$ and q is within the range of $-30°$ and $30°$. This threshold is relatively easy to achieve, which may have limited the extent to which our method could demonstrate its full potential.

Table 1. Quantitative comparison on SQA3D [4].

Method	Acc@0.5m	Acc@1.0m	Acc@15°	Acc@30°
SQA3D [4]	14.60	34.21	22.39	42.28
ours	**34.41**	**55.27**	**40.07**	**44.98**

References

[1] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *European Conference on Computer Vision (ECCV)*, 2022.

[2] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[3] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

[4] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022.

[5] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.