3D Visual Grounding with Graph and Attention

Tao Gu (tao.gu@tum.de) Yue Chen (yue.chen@tum.de)
Advisor: Dave Zhenyu Chen

Motivation

With the rapid development of 3D sensors and 3D representation, 3D Visual Grounding, which tries to segment out target objects in a point cloud using a linguistic description, has emerged as a new topic in the fields of 3D computer vision and natural language processing.

However, there are several obstacles. Unlike 2D images, 3D point clouds are more likely to be unordered and sparse, making it more challenging to locate objects and capture their relations. It's also difficult to extract precise object relations from descriptions and combine them with detected objects.

Goal

Our project aims to make full use of the advantage of graph and attention to increase the accuracy of localizing current objects, solving the problem of insufficient connections between instance proposals.

Dataset

We employed the **ScanRefer** dataset for training and evaluation. It contains 51,583 descriptions of 11,046 objects from 800 ScanNet scenes. Our method only takes coordinates (XYZ) and color (RGB) information as PointCloud input.

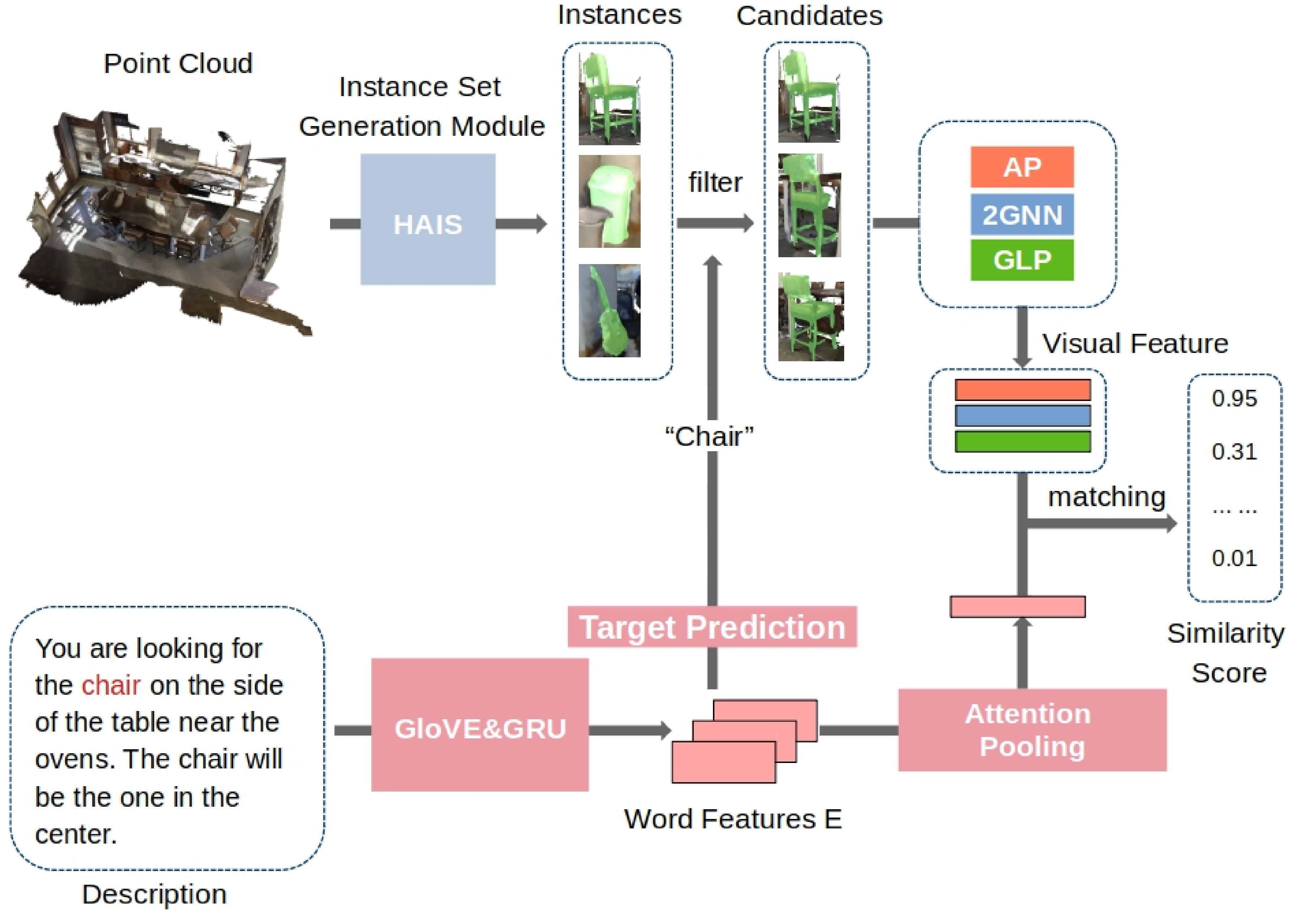
The small office chair. The The chair is in the corner by the table.

there is a black arm chair. placed in the corner of the office.

there is a black arm chair.placed next to another same chair.

...

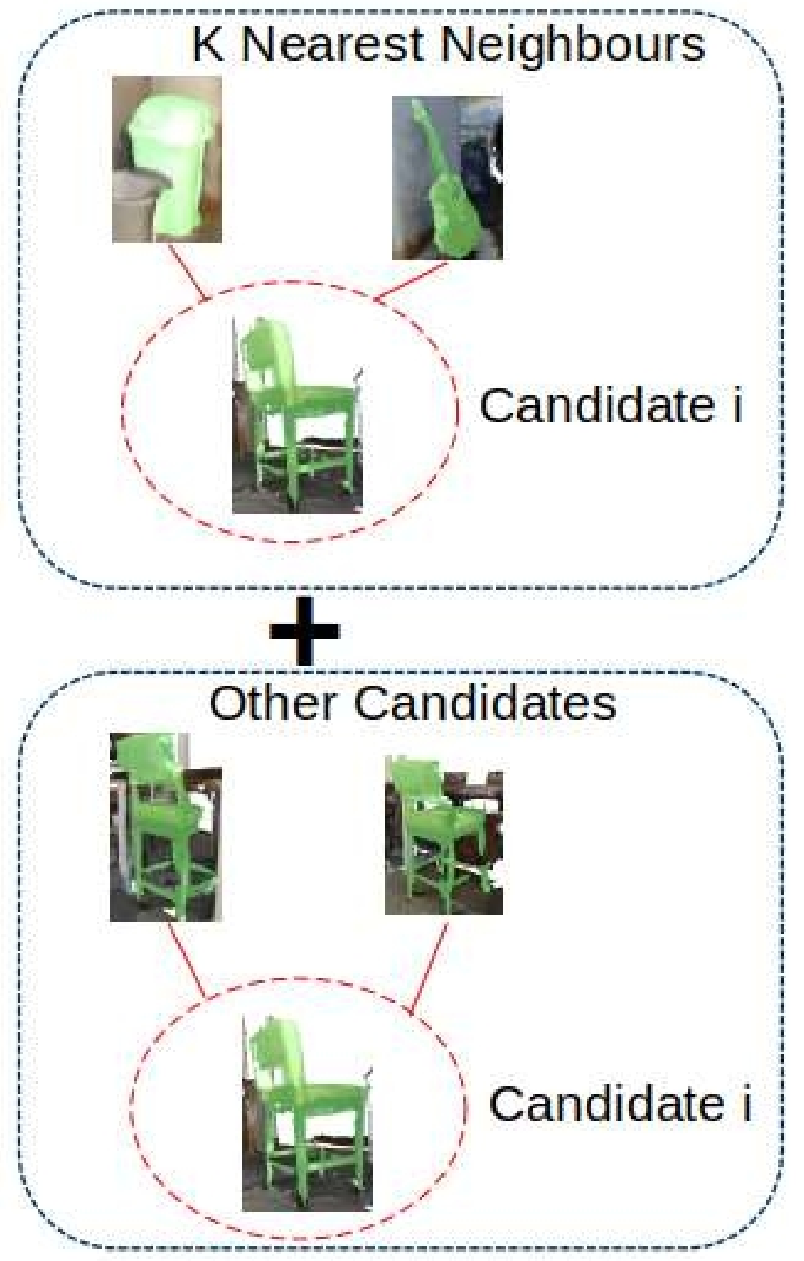
Architecture of HAIS_2GNN



Contribution

Our model **HAIS_2GNN** is developed based on InstanceRefer. Highlights of our model are:

- **HAIS**
  - It is an efficient framework for point cloud instance segmentation.
  - We applied the official pretrained model.
- **2GNN**
  - The first GNN captures the relation between each candidate and its k nearest neighbors.
  - The second GNN passes message among all candidates in a scene.



Result

Method	Unique Acc@0.5	Multiple Acc@0.5	Overall Acc@0.5
Validation results			
ScanRefer	43.31	19.75	24.32
InstanceRefer (baseline)	64.99	22.93	31.09
InstanceRefer w/ MAT (paper)	66.83	24.77	32.93
HAIS_2GNN (ours)	**70.24**	**24.85**	**33.66**

Our model achieves remarkable results, where:

- Our model **outperforms** the original InstanceRefer
- Only **a few parameters** are added in comparison to the baseline.
- There is **more than 8% improvement** in both "Unique" and "Multiple".
- It is feasible to incorporate co-attention fusion (MAT) to our model.

Conclusion

We proposed an improved framework named HAIS_2GNN based on the InstanceRefer via a more powerful instance set generation module HAIS and an enhanced relation module with two graph neural networks. Our model performs remarkable results with an accuracy of 33.66 in regards to the "overall" scene.

Future Work

For future work, we may attempt to improve our model by implementing a feasible attention-based fusion model or incorporating the co-attention fusion model from the InstanceRefer. Also, we would try to train an end-to-end model using unfrozen HAIS. Finally, we will submit our model to the benchmark.