3D Visual Grounding with Transformers - RefNetV2

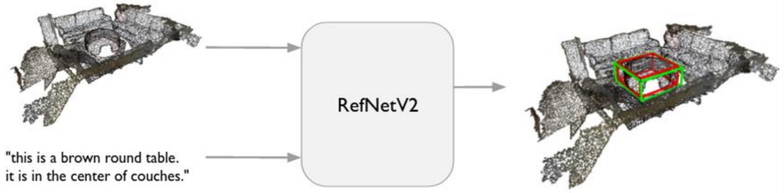
Philipp Foth[1], Bastian Wittmann[1], supervised by Dave Zhenyu Chen[1]
[1] Technical University of Munich, {philipp.foth, bastian.wittmann}@tum.de

Introduction

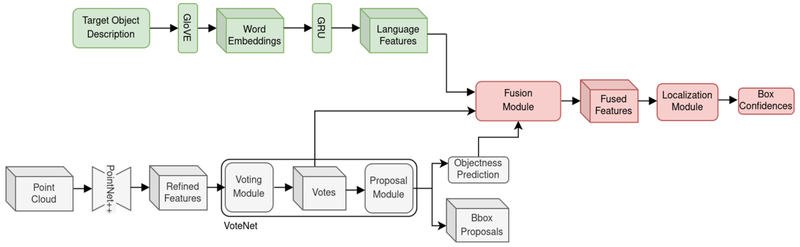
In visual grounding, an object is localized based on a natural language query. It is common to solve this task in two stages by first detecting all objects in the scene and then using the language component to choose the correct object proposal. The object detection in the first step plays a crucial role in the quality of the final output. For this reason, in this work we improve the visual grounding results of an existing method by extending it with a better object detector.

3D Visual Grounding

The task of 3D visual grounding is to locate an object in a 3D scene, represented for example by a point cloud, given a natural language description of that object..

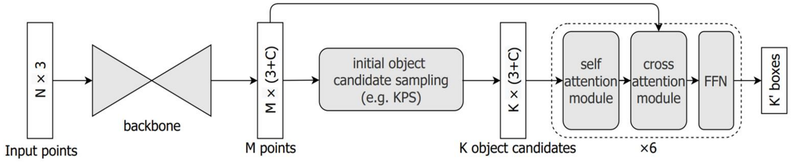


The baseline method ScanRefer first detects the objects in the scene with a VoteNet object detection module. Subsequently, the language description input is processed in a GRU RNN. The object and language features are fused together and finally used to predict the correct reference object.



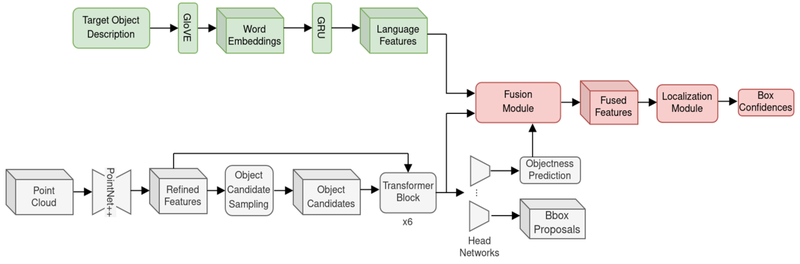
3D Object Detection

Recently, a transformer-based object detector has been shown to significantly outperform VoteNet in 3D object detection. In the transformer decoder, object candidates can aggregate information from other object candidates via a self-attention module and from all the points of the point cloud via a cross-attention module.



RefNetV2

We replaced the VoteNet detection module in the ScanRefer method with the transformer-based method shown above:



We use variations of this architecture for our evaluation and ablation studies.

Evaluation

We compare RefNetV2 with other recent visual grounding methods from the literature on the validation data of the ScanRefer dataset. Only the 3D coordinates are used as input features.

	Unique Acc		Multiple Acc		Overall Acc	
	@0.25IoU	@0.5IoU	@0.25IoU	@0.5IoU	@0.25IoU	@0.5IoU
ScanRefer	63.98	43.57	29.28	18.99	36.01	23.76
InstanceRefer	74.91	**64.23**	27.93	21.82	37.04	27.05
FFL-3DOG	-	64.04	-	24.13	-	32.47
Ours	**75.23**	64.17	**35.66**	**27.33**	**43.34**	**34.48**
Ours vs. ScanRefer	+11.3	+20.6	+6.4	+8.3	+7.3	+10.7

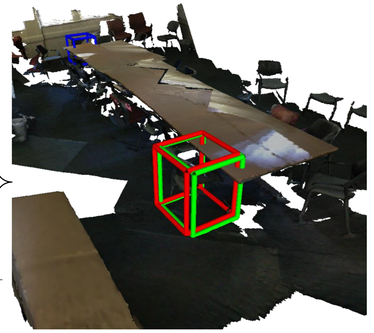
When visualizing the output bounding boxes of our method and comparing them to the output of the ScanRefer baseline, we observe that not only do the correctly matched boxes overlap better with the ground truth, but also the correct box is localized more often.

"the chair is the northeastern-most one next to the table. the chair is gray and has four legs."

"this is a black office chair. it's on the corner of the table **closest to the whiteboard**."

"the chair is the northwestern-most one on the right side of the table. the chair is gray and has four legs."

"the black chair, it was placed near a brown table, positioned right on top of the table, **near the wall**. to the right was a white table."



IoUs	desc. 1	desc. 2	desc. 3	desc.4
RefNetV2	0.00	**76.22**	0.00	**75.52**
ScanRefer	0.00	0.00	0.00	0.00

Future Work

An interesting future direction would be to incorporate the language input in the transformer decoder module, allowing object features to attend to certain language embeddings in the feature refinement steps.