3D Visual Grounding with Transformers

S. Frisch, F. Stilz

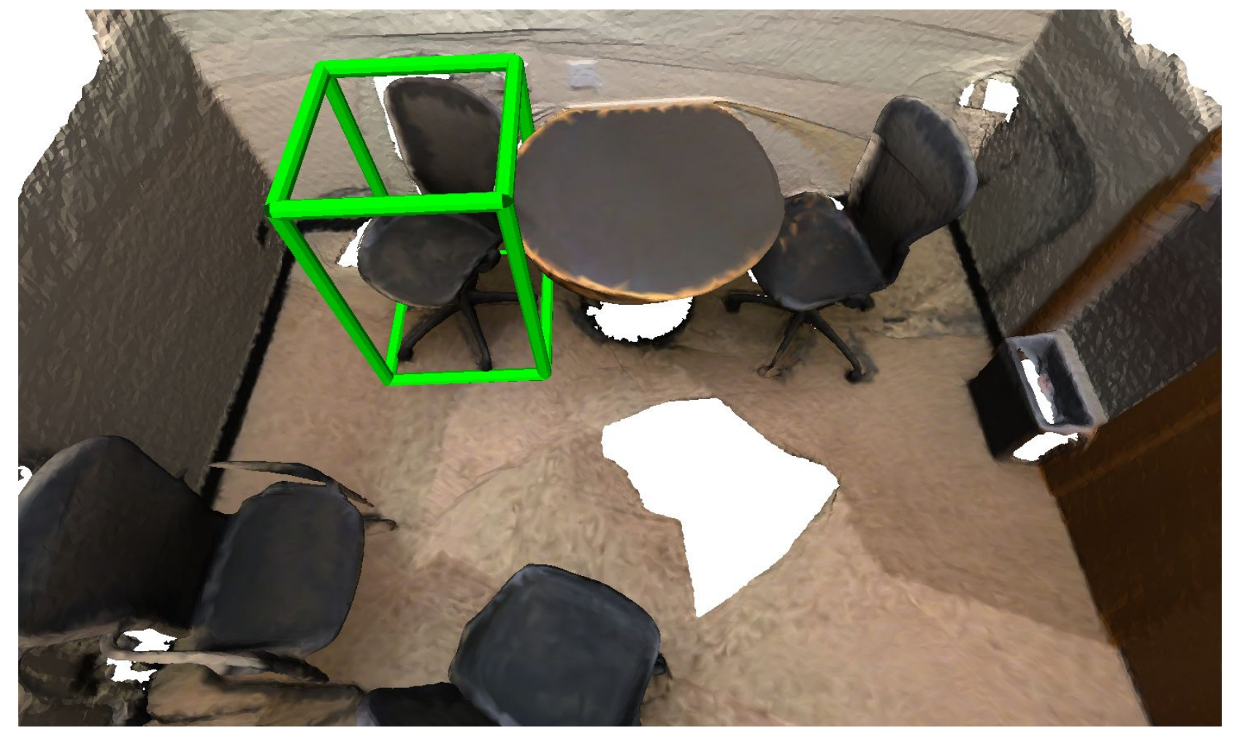
Technical University of Munich

Introduction

Inputs:

There is an outlet in the wall next to the black office chair. The office chair is next to a round table.

Output:

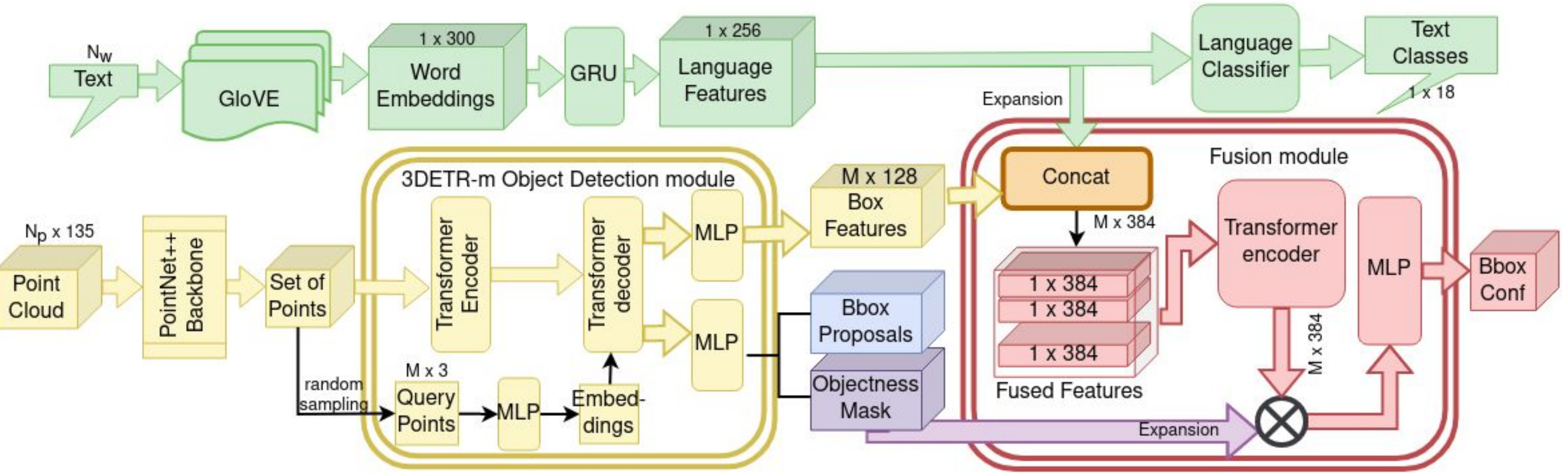


3D visual grounding is the task of localizing a target object in a 3D scene given a natural language description. The task involves 3D object detection, natural language encoding and the final fusion and localization of the target object.

Idea

We use a transformers-based architecture to fully utilize the contextual clues. Since transformers can naturally operate on variable sized inputs such as point cloud and encode long-range contexts that enrich the visual grounding task.

For 3D object detection and feature generation the **3DETR-m** transformer introduced into the ScanRefer architecture. Additionally, we used a **vanilla transformer encoder** for the matching of the textual and visual features since it is uniquely suited to capture the relevant context. We also experimented with replacing the GRU model with a transformer language encoder **BERT**.



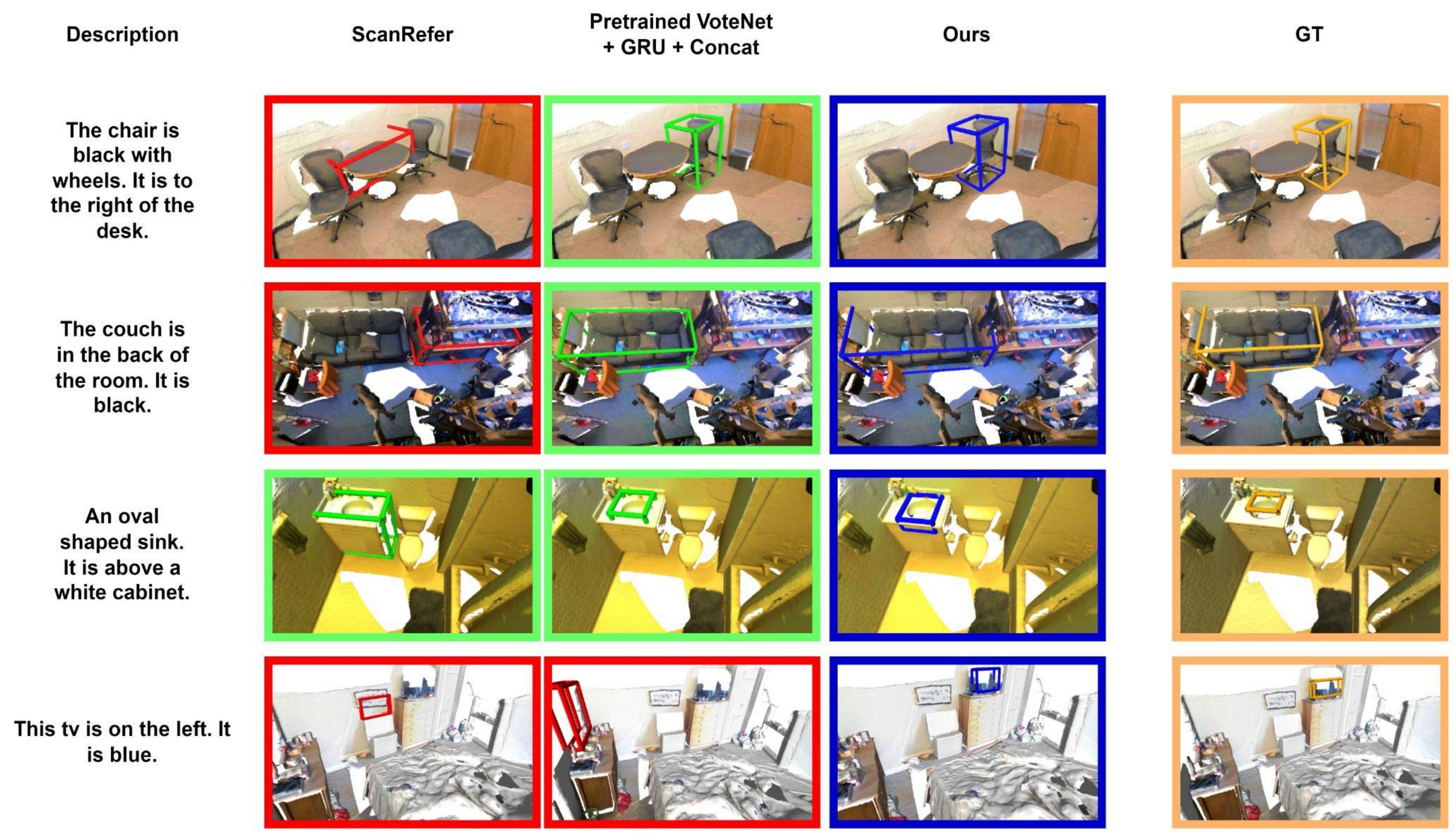
Ablation Study

Model	Acc@0.25	Acc@0.5
pretrained VoteNet + GRU + Concat	**37.11**	25.21
pretrained 3DETR-m + GRU + Concat	35.00	**25.50**

Model	Acc@0.25	Acc@0.5
ScanRefer (VoteNet + **GRU** + Concat)	**35.66**	**22.01**
VoteNet + **BERT Layer 12** + Concat	34.22	21.10

Model	Acc@0.25	Acc@0.5
pretrained 3DETR-m + GRU + **Concat**	35.00	25.50
pretrained 3DETR-m + GRU + **vTransformer Layer 5**	**37.08**	**26.56**

Qualitative Results



The chair is black with wheels. It is to the right of the desk.

The couch is in the back of the room. It is black.

An oval shaped sink. It is above a white cabinet.

This tv is on the left. It is blue.

Model	Acc@0.25	Acc@0.5	Duration
ScanRefer (VoteNet + GRU + Concat)	**37.05**	23.93	25h
Pretrained VoteNet + GRU + Concat	**37.11**	25.21	**4h 17min**
Ours (Pretrained 3DETR-m + GRU + vTransformer)	37.08	**26.56**	9h 20min

Check out our paper and project page for more results and comparisons!
https://github.com/flo-stilz/3D-Visual-Grounding-with-Transformers