Technische Universität München

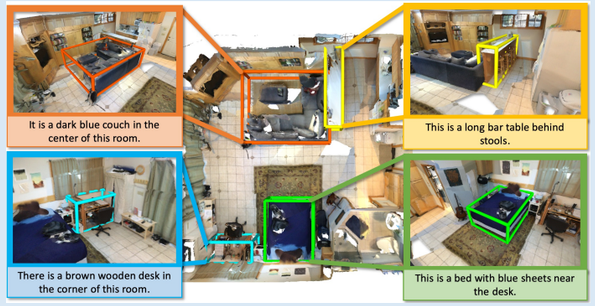
Joint Architecture for 3D Vision Language

Group Member: Robin Borth, Yaomengxi Han Supervisor: Dave Zhenyu Chen

code available

Introduction to tasks and datasets

- **Task**
  - 3D Visual Grounding: given **3D point cloud** and **natural language query** as input, localize the requested object.
  - 3D Dense Captioning: given a **3D point cloud** as input, **detect** all objects and **generate captions** for them.
- **Dataset**
  - **ScanNet**: contains **2.5M** views in more than **1500** scans with 3D camera poses, surface reconstructions, and instance-level semantic segmentations.
  - **ScanRefer:** contains **51,583** descriptions of **11,046** objects from **800** ScanNet scenes.

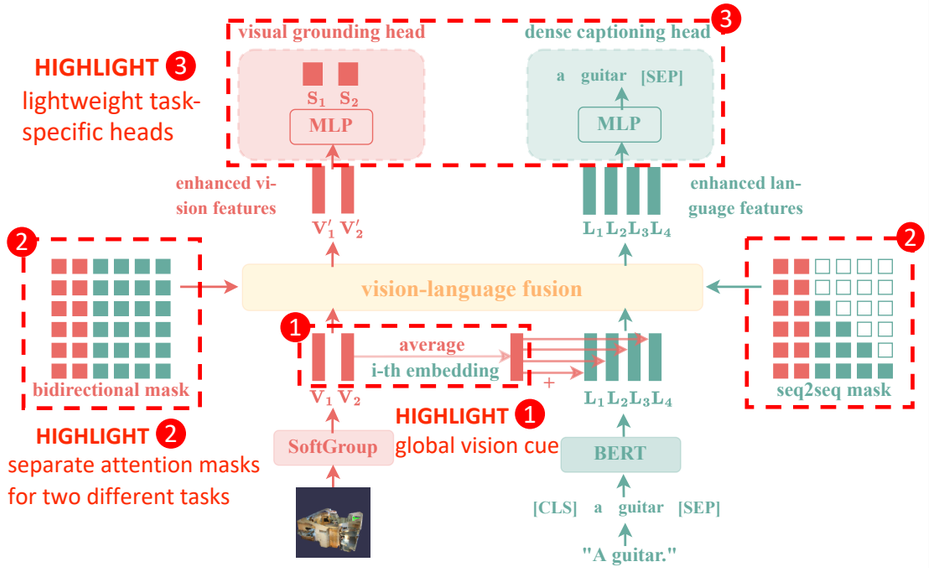


Selected objects and associated descriptions from one scene in ScanRefer Dataset.

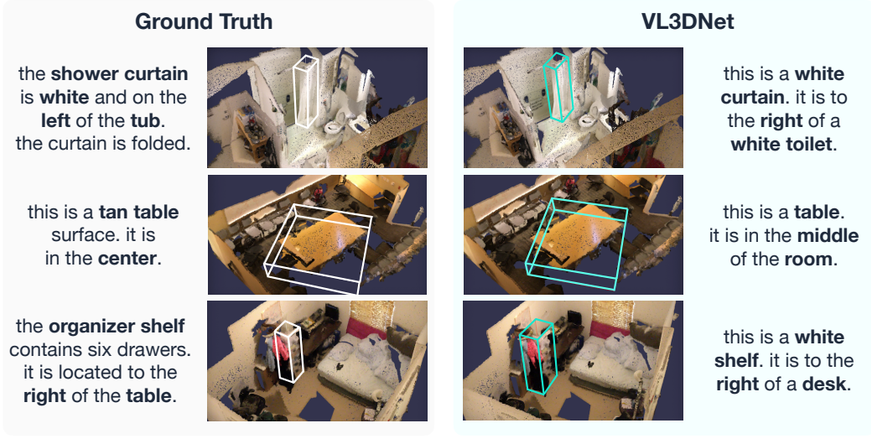
Motivation

- The **multimodal input** motivates us to…
  fuse vision-language information
- **Complementary nature** of the two tasks motivates us to…
  jointly train these two tasks
- The great impact of **strong detectors** motivates us to…
  use SoftGroup as detection backbone

VL3DNet architecture



Qualitative results and analysis



*Quantitative results

Methods	Visual Grounding: accuracy@0.5IoU			Dense Captioning: captioning F1-score	
	unique	multiple	overall	CIDEr	METEOR
**baseline	0.4352	0.2097	0.2603	0.0849	0.0492
D3Net	**0.6843**	0.3074	0.3919	0.2088	0.1022
3DJCG	0.6059	0.3117	0.3776	0.1918	0.1013
VL3DNet	0.6633	**0.3125**	**0.4172**	**0.2849**	**0.2003**

* We select a few important metrics due to the page limit.
** The baseline methods are ScanRefer for visual grounding and Scan2Cap for dense captioning.

Ablation studies

Grounding Only vs. Joint Training

training	accuracy@0.25IoU			accuracy@0.5IoU		
	unique	multiple	overall	unique	multiple	overall
vg	0.7174	0.3802	0.4829	0.6553	0.2980	0.4040
joint	**0.7291**	**0.3926**	**0.4940**	**0.6633**	**0.3125**	**0.4172**

Captioning Only vs. Joint Training

training	captioning F1-score			
	CIDEr	BLEU-4	Rouge-L	METEOR
dc	0.2800	**0.0702**	0.2091	0.1995
joint	**0.2849**	0.0698	**0.2115**	**0.2003**