Lang3DSG

Language-based contrastive pre-training for 3D Scene Graph prediction

BOSCH [1]

universität ulm [2]

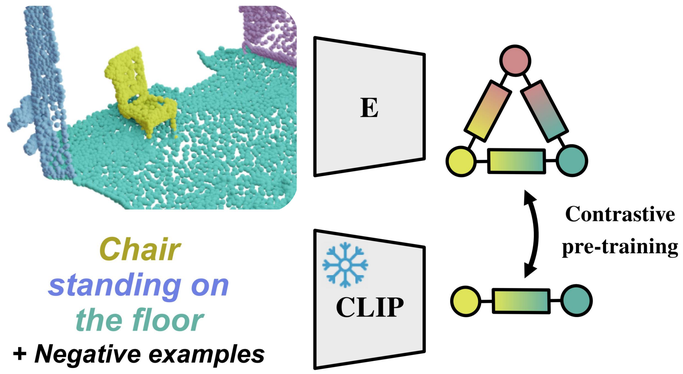
TECHNISCHE UNIVERSITÄT WIEN [3]

Sebastian Koch[1,2] Pedro Hermosilla[3] Narunas Vaskevicius[1] Mirco Colosi[1] Timo Ropinski[2]

3DV 2024

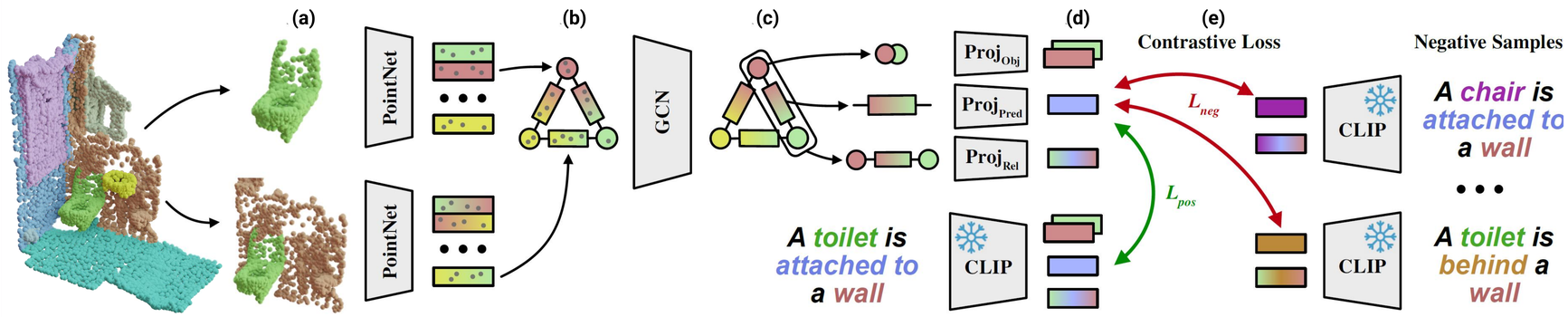
1. Introduction

Key Insight: Language and Scene Graphs are very similar



Goal: Improved 3D scene graph prediction by language supervised pre-training

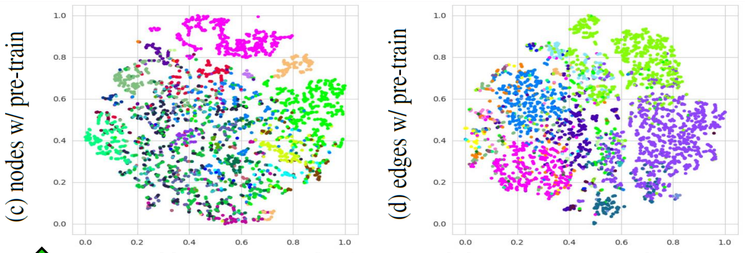
2. Method



Construct a latent scene graph from a 3D point cloud. Then align the 3D scene graph embedding with the well-structured language model embedding (CLIP/BERT) using a contrastive loss.

$$\mathcal{L}_{pos} = \sum_{i=1}^{N} \frac{1}{|K|} \sum_{j \in K} 1 - \cos(f_i, f^t_{h(j)}) \qquad \mathcal{L}_{neg} = \sum_{i=1}^{N} \frac{1}{|M|} \sum_{j \in M} \max\left(0, \cos(f_i, f^t_{h(j)}) - \tau\right)$$

3. Object & Predicate Embedding



Structured latent space for improved downstream prediction

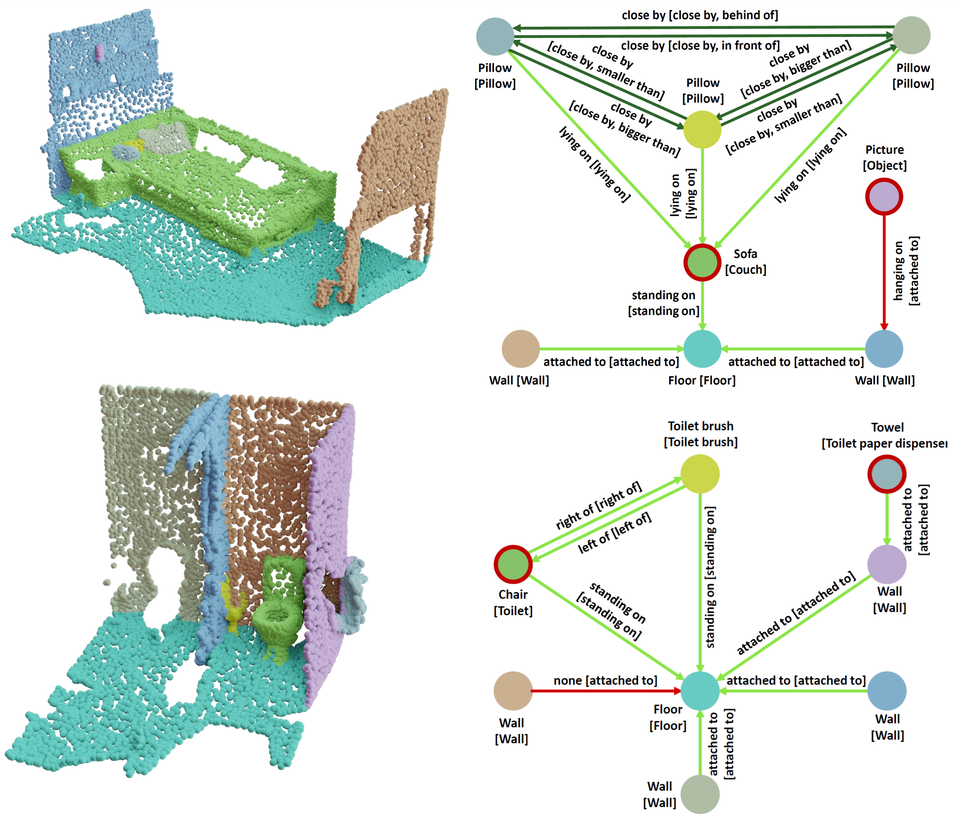
4. Quantitative Results

	Object		Predicate		Relationships	
	R@5	R@10	R@3	R@5	R@50	R@100
3DSSG	0.68	0.78	0.89	0.93	0.40	0.66
SGFN	0.70	0.80	**0.97**	**0.99**	0.85	0.87
Lang3DSG	**0.77**	**0.84**	**0.97**	**0.99**	**0.87**	**0.89**

	Pre-train		Object		Predicate	
	PCL	SG	R@5	mR@5	R@3	mR@3
Ours (no pre-train)			0.63	0.30	0.94	0.57
STRL	✓		0.75	0.35	0.94	0.50
DepthContrast	✓		0.77	0.36	0.94	0.51
Ours (no GCN)		✓	0.74	0.37	0.94	0.60
Ours		✓	**0.77**	**0.43**	**0.96**	**0.67**

Scene Graph predictions requires graph-based pre-training

5. 3D Semantic Scene Graph Predictions



6. Zero-shot room type classification