Object-Centric Representation Learning for Enhanced semantic 3D scene graph Prediction

KunHo Heo*[1], GiHyeon Kim*[1], SuYeon Kim[1], MyeongAh Cho[1]
[1]Department of Software Convergence, KyungHee University
*Equal Contribution

Project Page: https://visualsciencelab-khu.github.io/OCRL-3DSSG/

Arxiv Link: https://arxiv.org/pdf/2510.04714

Introductions

- 3D (Semantic) Scene Graph(3DSG) prediction aims to build the graph representation on given 3D scene (point cloud, mesh, etc…)
- Scene graph estimates the semantic labels of **vertices(objects)** and **their edge(relationship)**.

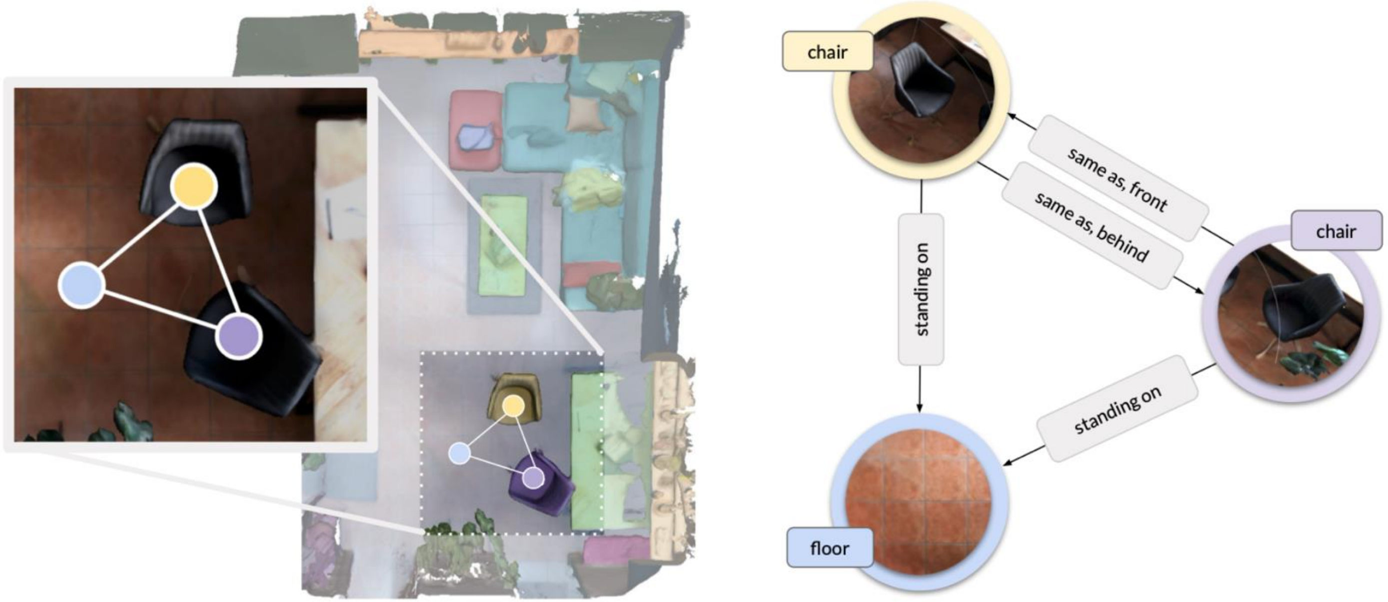
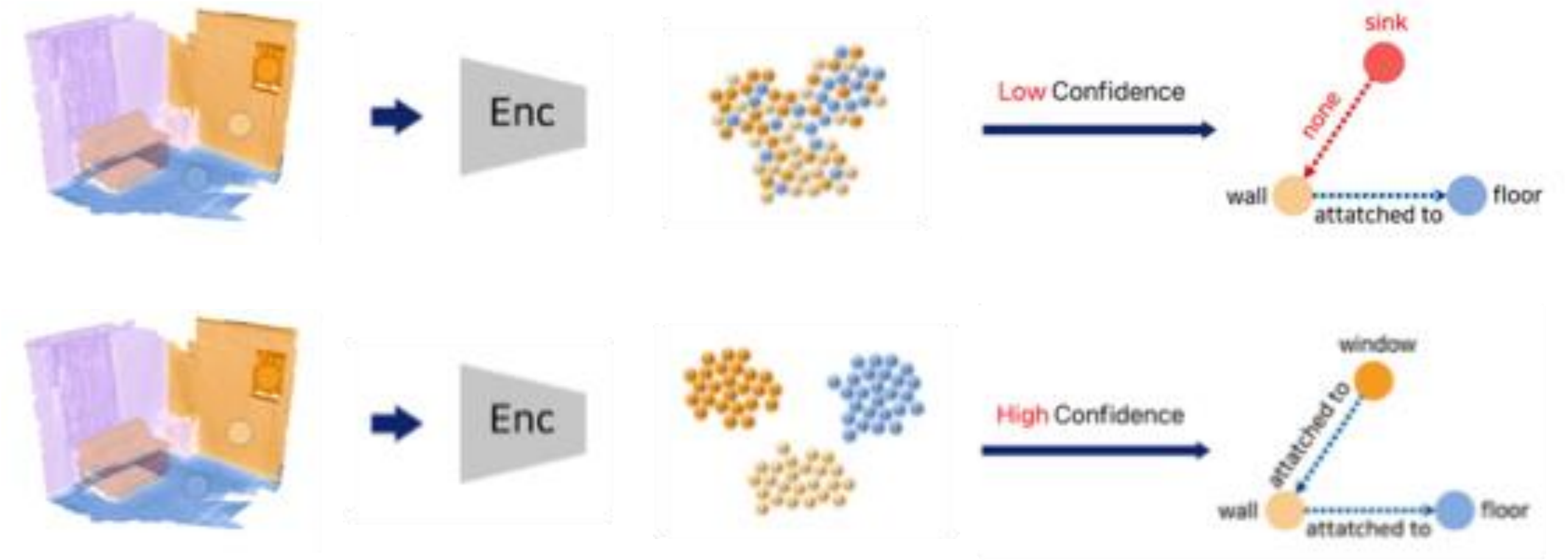


Fig. Overview of 3D semantic scene graph prediction task

Motivations & Observations

- As empirical observations, our study demonstrated that core bottleneck of 3D scene graph prediction is ‘object representation’.
- The more accurate and discriminate object features are, the better performance of 3D scene graph come.



Motivations & Observations

- Through our empirical studies, we hypothesize that performance of 3D scene graph will rely – explicitly or implicitly – on accuracy of classification and confidence of estimation.
- As our hypothesis, we can probabilistically formulate the prediction model like:

$$P(e_{ij}|z_i, z_j) = \sum_{o'_i, o'_j \in \mathcal{O}} P(e_{ij}|o'_i, o'_j)P(o'_i|z_i)P(o'_j|z_j)$$

Model	Obj. ✓ Sub. ✓	Obj. ✓/✗ Sub. ✗/✓	Obj. ✗ Sub. ✗
SGPN [45]	8%	12%	18%
SGFN [52]	8%	12%	20%
VL-SAT [48]	8%	13%	19%

Tab.1. Misclassification rate of predicate respect to the object misclassification

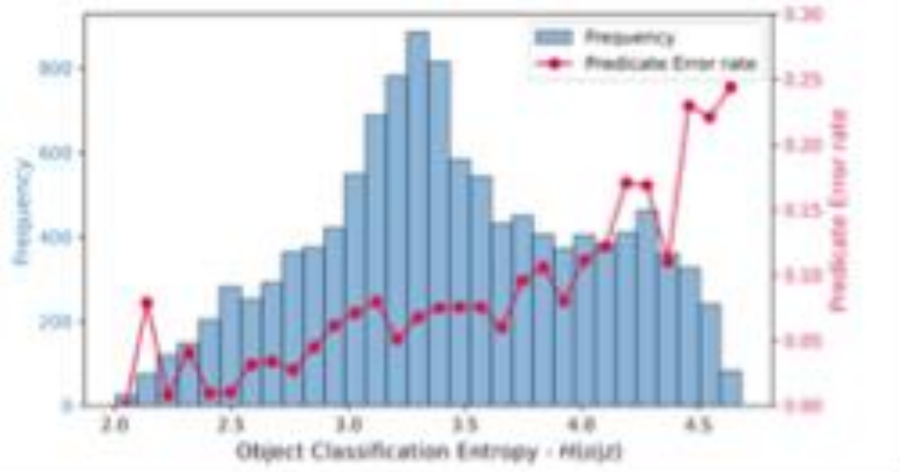
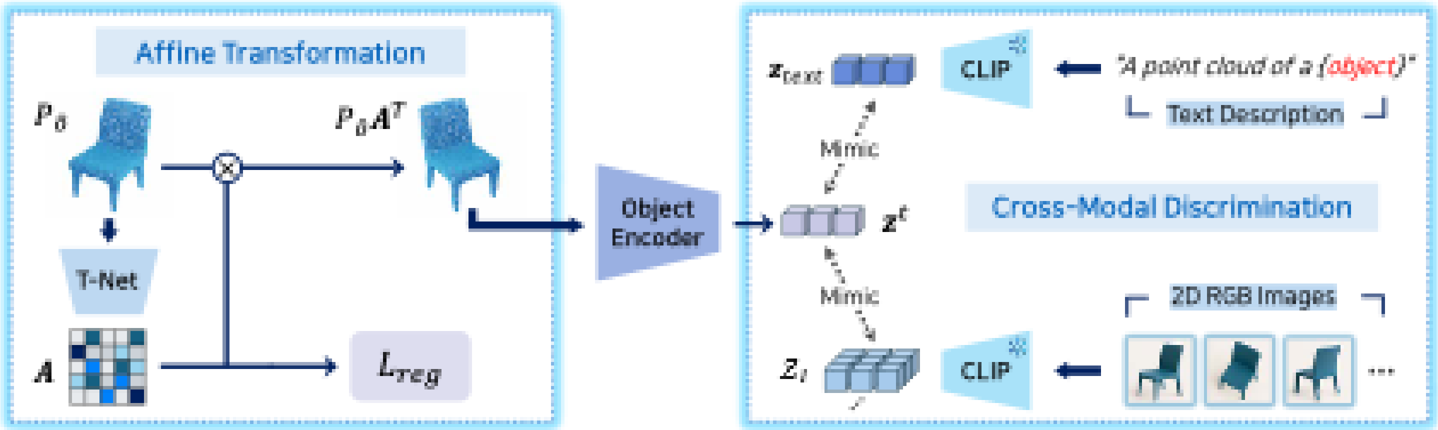


Fig. Misclassification rate of predicate respect to the object classification confidence

Methods

- We configured our proposed method as two stage: (1) Discriminative object feature pretraining, (2) training GNN model leveraging discriminative object representation.
- As Fig.2, we used CLIP text/image features to divide object features with supervised contrastive learning.

$$\mathcal{L}_i^{\text{text}} = -\log \frac{\exp s(z_i^t, z_{\text{text}}^i)/\tau}{\sum_{r\in \mathcal{N}(i)} \exp s(z_i^t, z_{\text{text}}^r)/\tau}$$



$$\mathcal{L}_i^{\text{visual}} = \frac{1}{|P(i)|}\sum_{p\in\mathcal{P}(i)}\sum_{z_+\in Z_i^p} -\log \frac{\exp s(z_i^t, z_+)/\tau}{\sum_{r\in \mathcal{N}(i)}\sum_{z_-\in Z_i^r} \exp s(z_i^t, z_-)/\tau}$$

Fig. Architecture of proposed object feature encoder pre-trainer

Methods

- Given well-defined feature space, we focused on geometric information which cannot be contained solely in object feature.
- We propose Local/Global Spatial Enhancement and Bidirectional Edge Gating modules to aid scene graph prediction.

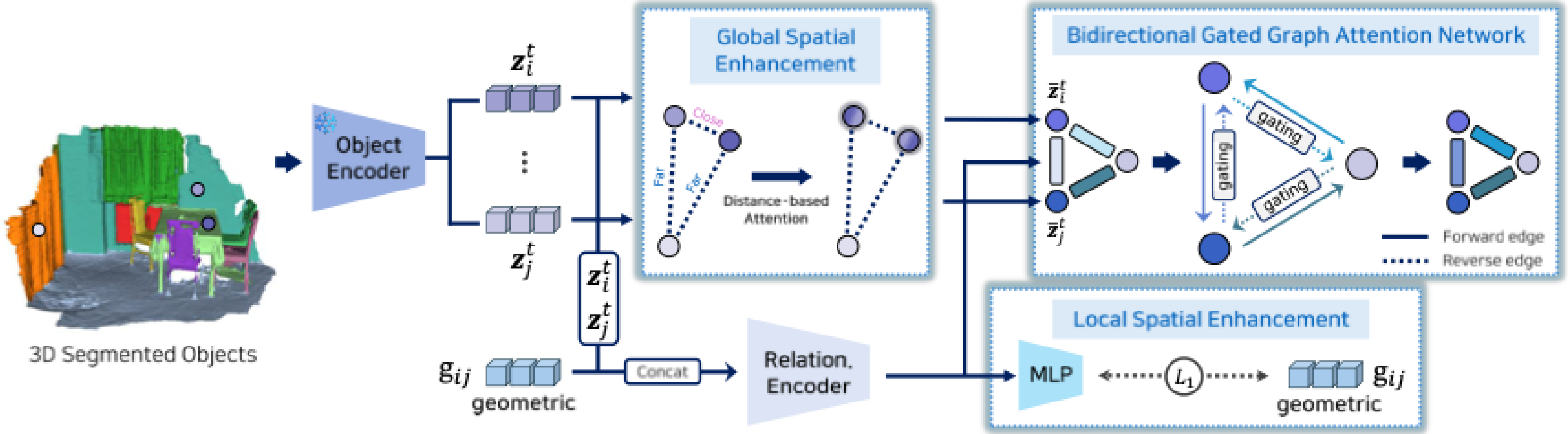


Fig. Architecture of proposed 3D scene graph prediction model

Experiments

- Our model outperformed previous 3DSG model baselines.
- Performance of object as well as predicate largely enhanced as our hypothesis

Model	Object		Predicate		Triplet	
	R@1	R@5	R@1	R@3	R@50	R@100
SGPN [45]	49.46	73.99	86.92	94.76	85.38	88.59
SGFN [52]	53.36	76.88	89.00	97.71	88.59	91.14
VL-SAT [48]	55.93	78.06	89.81	98.46	89.35	92.20
Ours	**59.53**	**81.20**	**91.27**	**98.48**	**91.40**	**93.80**

Model	SGCls (w/ GC)			PredCls (w/ GC)			SGCls (w/o GC)			PredCls (w/o GC)		
	R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100
SGPN [45]	27.0	28.8	29.0	51.9	58.0	58.5	28.2	32.6	35.3	54.5	70.1	82.4
Zhang et al. [63]	28.5	30.0	30.1	59.3	65.0	65.3	29.8	34.3	37.0	62.2	78.4	88.3
SGFN [52]	29.5	31.2	31.2	65.9	78.8	79.6	31.9	39.3	45.0	68.9	82.8	91.2
VL-SAT [48]	32.0	33.5	33.7	67.8	79.9	80.8	33.8	41.3	47.0	70.5	85.0	92.5
Ours	**36.1**	**37.7**	**37.8**	**70.2**	**82.0**	**82.6**	**38.1**	**46.1**	**52.5**	**73.3**	**87.8**	**94.6**

Tab.2&3. Overall performance of 3D scene graph prediction

Experiments

- To solid our claim, we adopted our object feature encoder into other baselines.
- As shown in Tab.4, the PredCls performance showed huge performance gain, which empirically supports our hypothesis.

Model	OFL (ours)	Object		Predicate		Triplet		SGCls		PredCls	
		R@1	R@5	R@1	R@3	R@50	R@100	R@20	R@50	R@20	R@50
SGPN [45]	✗	47.37	72.00	88.60	97.15	85.83	89.06	22.9	24.0	63.9	75.3
	✓	54.49	75.02	90.10	98.06	88.83	91.16	29.8	31.0	68.2	79.0
		+7.12%	+3.02%	+1.50%	+0.91%	+3.00%	+2.10%	+6.9%	+7.0%	+4.3%	+3.7%
SGFN [52]	✗	56.18	78.04	89.61	98.01	89.50	92.05	31.5	33.0	67.7	79.2
	✓	58.75	79.70	89.63	98.24	89.99	92.41	35.0	36.3	70.7	80.9
		+2.57%	+1.66%	+0.02%	+0.23%	+0.49%	+0.36%	+3.5%	+3.3%	+3.0%	+1.7%
VL-SAT [48]	✗	55.68	78.06	89.81	98.45	89.43	92.22	32.0	33.5	67.8	80.0
	✓	59.30	80.67	90.48	98.51	90.40	93.03	34.9	36.6	70.6	81.7
		+3.62%	+2.61%	+0.67%	+0.06%	+0.97%	+0.81%	+2.9%	+3.1%	+2.8%	+1.7%

Tab.4. Ablation studies of object feature encoder

Experiments

- Also, we proved that each of geometric modules plays important roles, which shows certain performance gain on Top-K mean Recall.
- We can infer that accurate predicate estimation requires discriminative object and proper information.
- Since predicate label in 3DSSG dataset is mostly spatial-relevant, we can achieved great performance with only two factors.

GSE	BEG	LSE	Object		Predicate		Triplet		SGCls		PredCls	
			R@1	mR@1	R@1	mR@1	R@50	mR@50	R@50	mR@50	R@50	mR@50
			58.02	20.77	90.55	50.36	90.19	61.79	43.8	36.5	85.7	68.3
✓			59.28	21.10	90.69	50.80	**91.51**	62.59	46.0	39.9	87.0	68.5
✓	✓		59.49	22.17	90.65	53.81	91.18	64.83	45.7	43.0	86.7	73.2
✓	✓	✓	**59.53**	**22.56**	**91.27**	**56.32**	91.40	**65.31**	**46.1**	**44.5**	**87.7**	**74.7**

Tab.5. Ablation studies of spatial modules

Summary

- Re-examined the importance of object representation in 3DSG prediction and its condition – discriminative feature space.
- Proposed simple yet efficient methodologies to improve performance of 3DSG, which overwhelms previous studies.
  - Object Feature Learning for discriminative object feature space using Supervised Contrastive Learning
  - Local Spatial Enhancement, novel auxiliary task to capture local geometric information
  - Global Spatial Enhancement to integrate contextual information among object instances
  - Bidirectional Edge Gating to regulate directional information between objects.
- Provided thorough theoretical and empirical analysis to describe detailed conditions of our findings.