JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

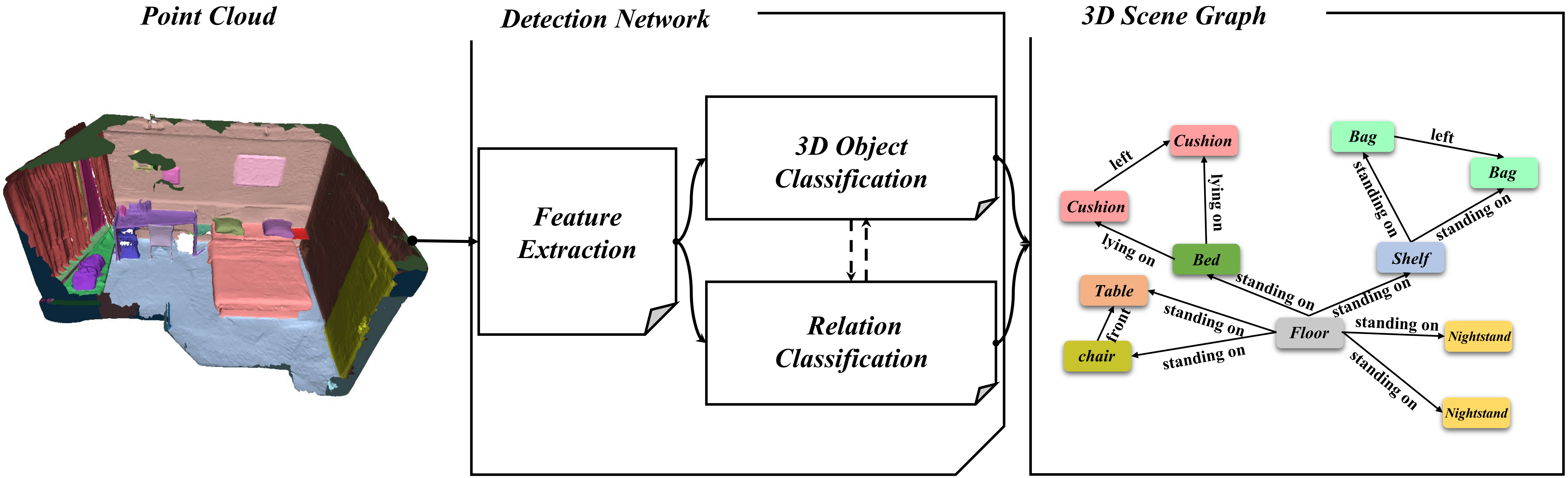
3D Spatial Multimodal Knowledge Accumulation for Scene Graph Prediction in Point Cloud

Mingtao Feng[1*] , Haoran Hou[1*] , Liang Zhang[1†] , Zijie Wu[2†],
Yulan Guo[3] , Ajmal Mian[4]

[1]Xidian University [2]Hunan University [3]Sun Yat-Sen University [4]The University of Western Australia

*Equal contribution
†Corresponding author

Problem Formulation



Point Cloud

Detection Network

Feature Extraction

3D Object Classification

Relation Classification

3D Scene Graph

1. Rich structural variations

2. Noisy, cluttered nature of real 3D scans

3. Uneven distribution over different objects and relationships

…

1.Inaccurate

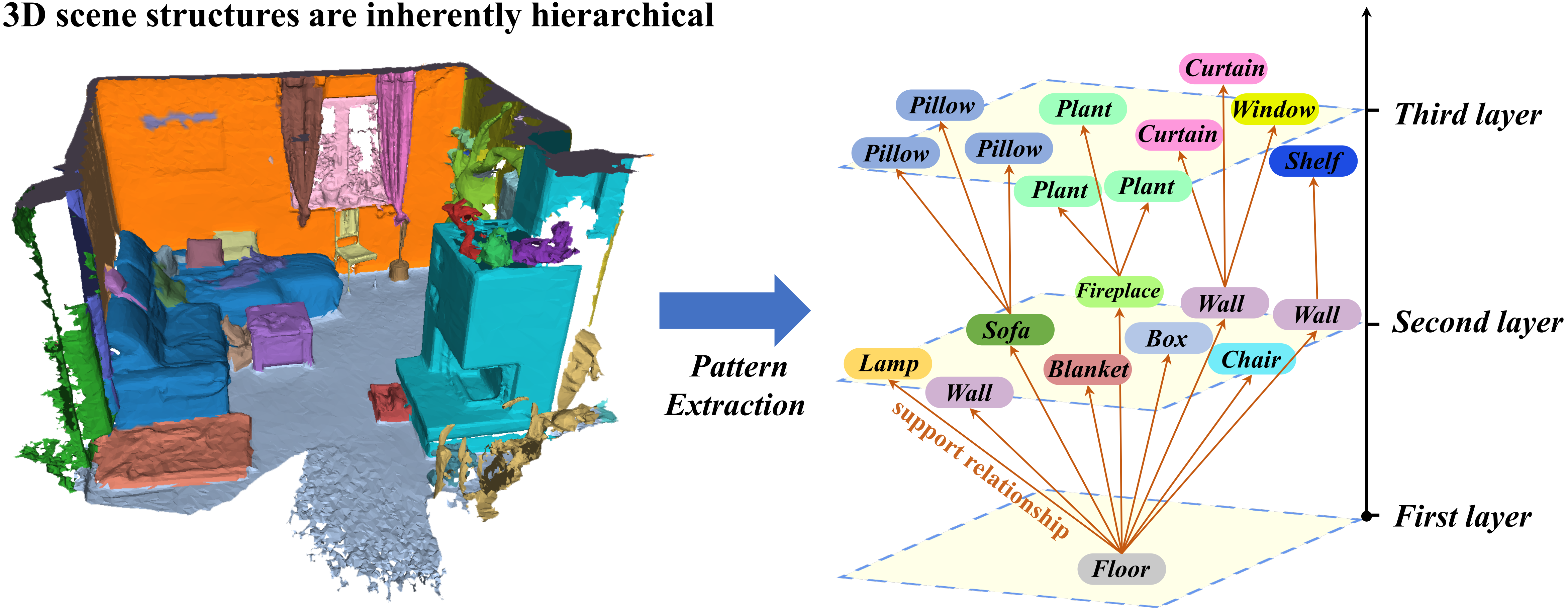
2.Lacks of robustness

3.Suffers from Long-tail problem

…

Motivation

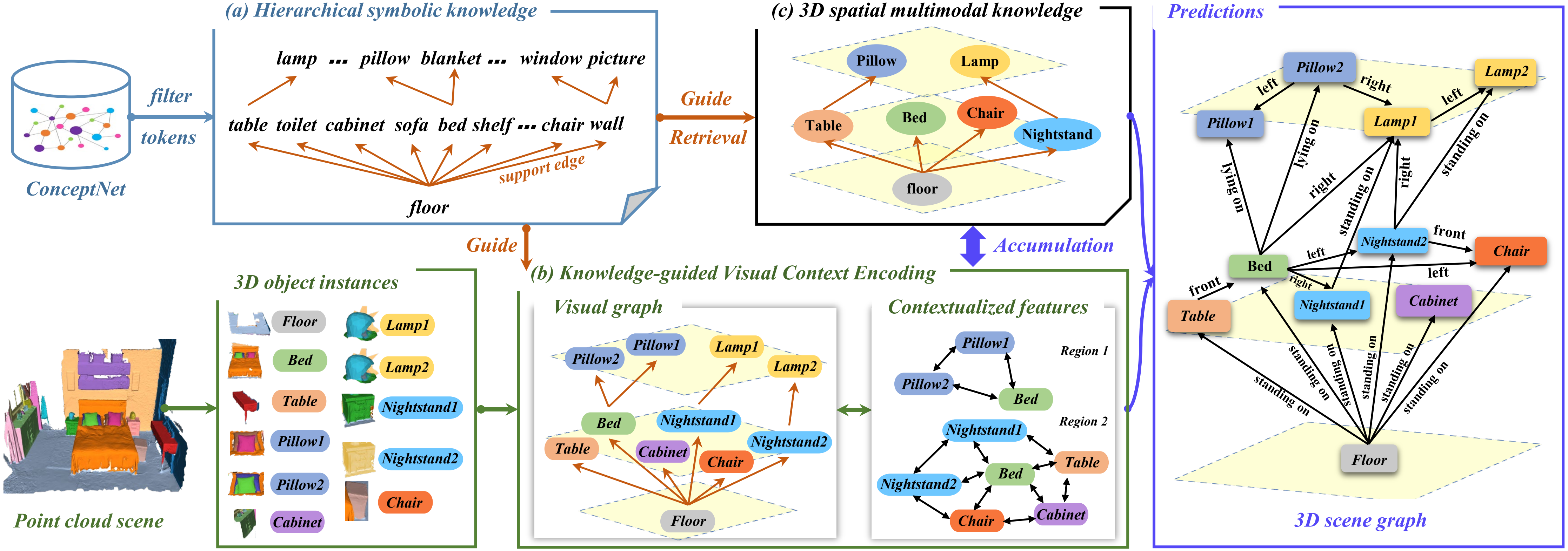
- **3D scene structures are inherently hierarchical**



- **How to extract the hierarchical structure patterns of any given 3D scene?**

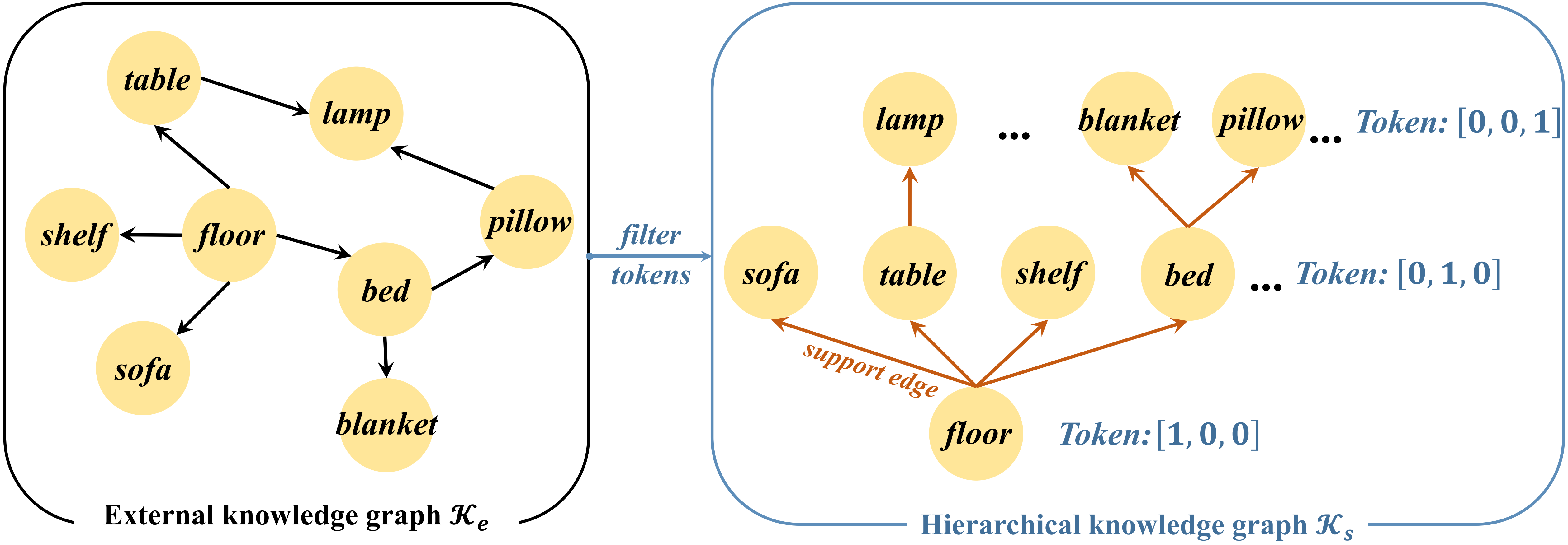
Our approach

Overview



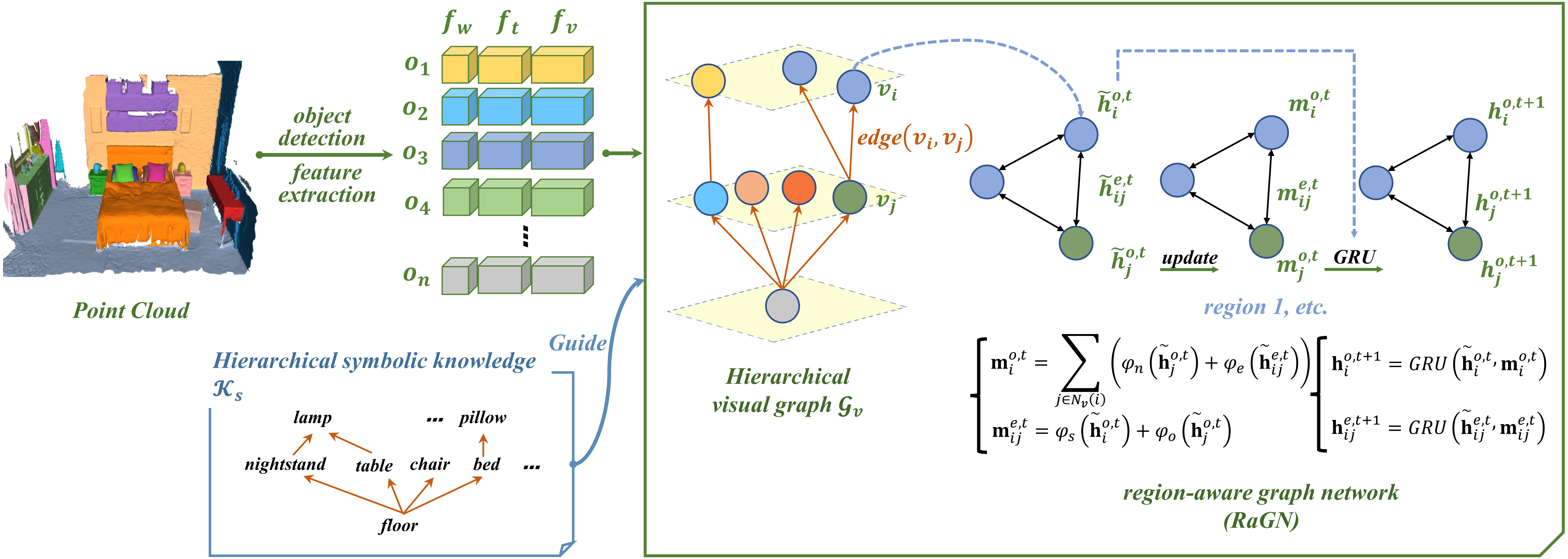
Our approach

(1) Hierarchical Symbolic Knowledge Initialization



Our approach

(2) Knowledge-guided Visual Context Encoding

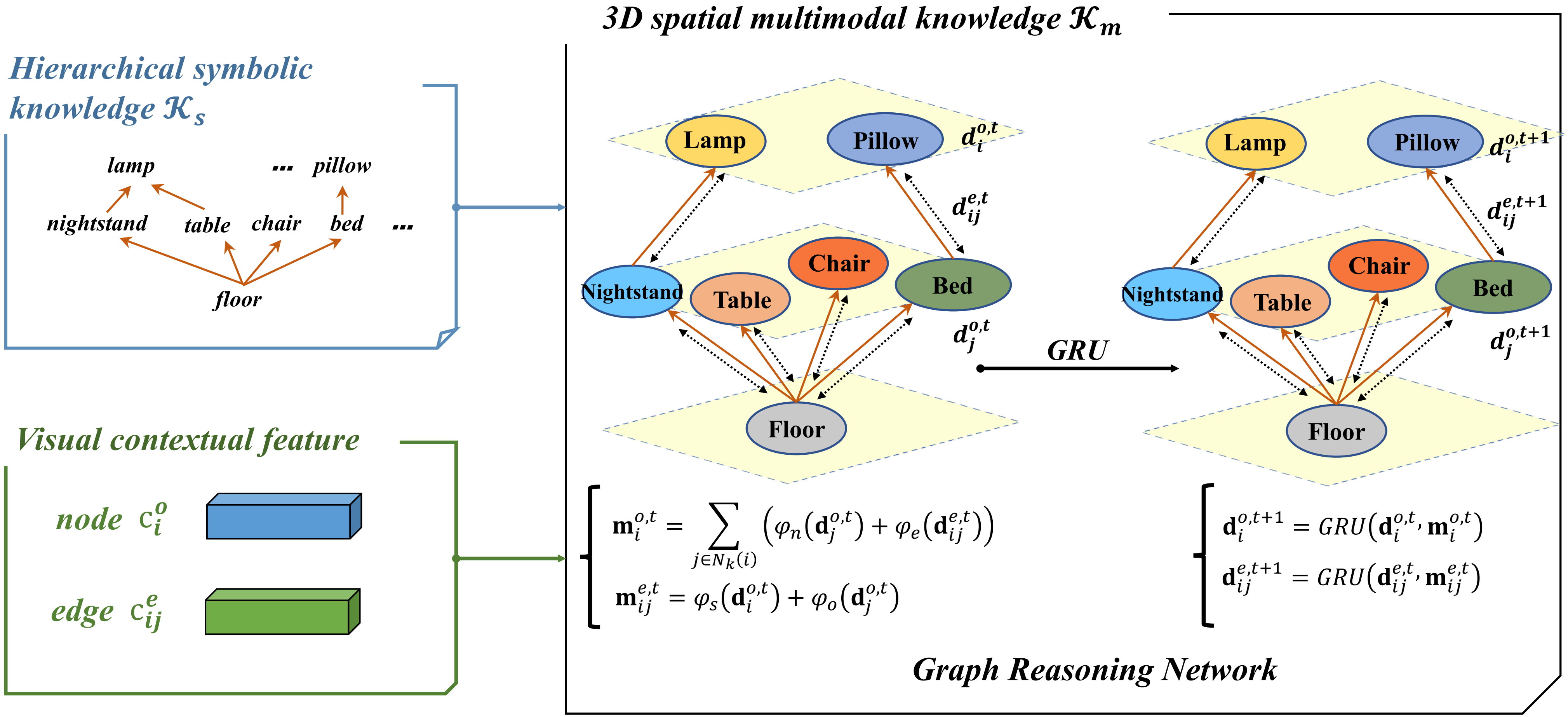


$$\begin{cases} \mathbf{m}_i^{o,t} = \sum_{j\in N_v(i)} \left(\varphi_n\left(\tilde{\mathbf{h}}_j^{o,t}\right) + \varphi_e\left(\tilde{\mathbf{h}}_{ij}^{e,t}\right)\right) \\ \mathbf{m}_{ij}^{e,t} = \varphi_s\left(\tilde{\mathbf{h}}_i^{o,t}\right) + \varphi_o\left(\tilde{\mathbf{h}}_j^{o,t}\right) \end{cases}$$

$$\begin{cases} \mathbf{h}_i^{o,t+1} = GRU\left(\tilde{\mathbf{h}}_i^{o,t}, \mathbf{m}_i^{o,t}\right) \\ \mathbf{h}_{ij}^{e,t+1} = GRU\left(\tilde{\mathbf{h}}_{ij}^{e,t}, \mathbf{m}_{ij}^{e,t}\right) \end{cases}$$

Our approach

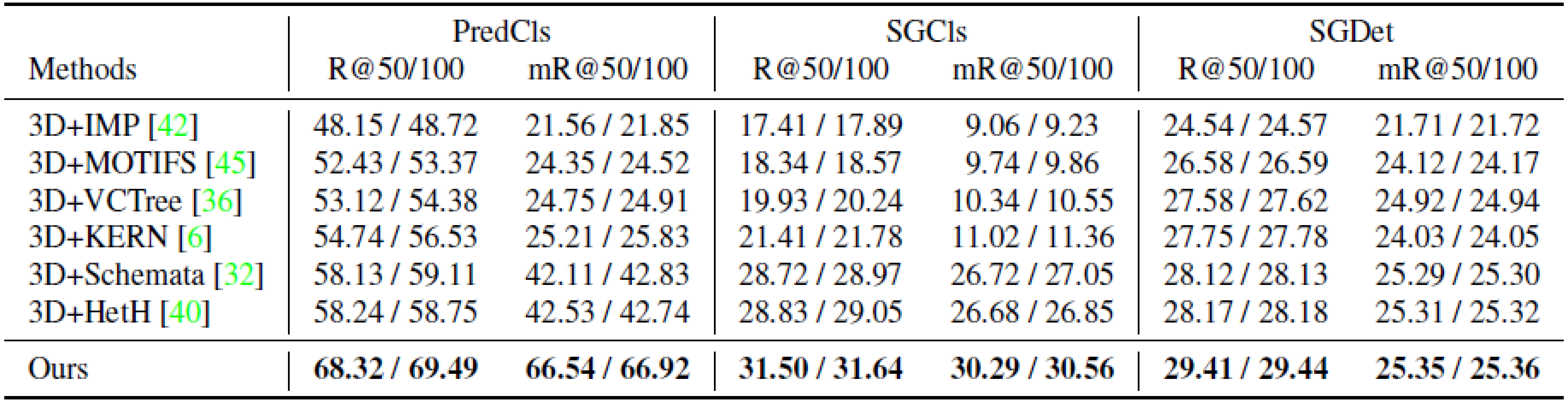
(3) Spatial Multimodal Knowledge Accumulation



$$\begin{cases} \mathbf{m}_i^{o,t} = \sum_{j \in N_k(i)} \left(\varphi_n(\mathbf{d}_j^{o,t}) + \varphi_e(\mathbf{d}_{ij}^{e,t}) \right) \\ \mathbf{m}_{ij}^{e,t} = \varphi_s(\mathbf{d}_i^{o,t}) + \varphi_o(\mathbf{d}_j^{o,t}) \end{cases}$$

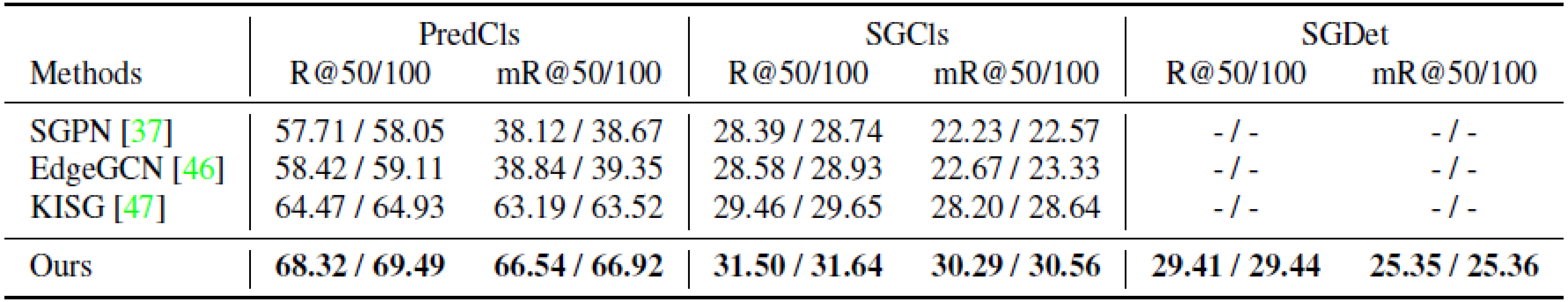
$$\begin{cases} \mathbf{d}_i^{o,t+1} = GRU(\mathbf{d}_i^{o,t}, \mathbf{m}_i^{o,t}) \\ \mathbf{d}_{ij}^{e,t+1} = GRU(\mathbf{d}_{ij}^{e,t}, \mathbf{m}_{ij}^{e,t}) \end{cases}$$

Experiment Results



Methods	PredCls		SGCls		SGDet	
	R@50/100	mR@50/100	R@50/100	mR@50/100	R@50/100	mR@50/100
3D+IMP [42]	48.15 / 48.72	21.56 / 21.85	17.41 / 17.89	9.06 / 9.23	24.54 / 24.57	21.71 / 21.72
3D+MOTIFS [45]	52.43 / 53.37	24.35 / 24.52	18.34 / 18.57	9.74 / 9.86	26.58 / 26.59	24.12 / 24.17
3D+VCTree [36]	53.12 / 54.38	24.75 / 24.91	19.93 / 20.24	10.34 / 10.55	27.58 / 27.62	24.92 / 24.94
3D+KERN [6]	54.74 / 56.53	25.21 / 25.83	21.41 / 21.78	11.02 / 11.36	27.75 / 27.78	24.03 / 24.05
3D+Schemata [32]	58.13 / 59.11	42.11 / 42.83	28.72 / 28.97	26.72 / 27.05	28.12 / 28.13	25.29 / 25.30
3D+HetH [40]	58.24 / 58.75	42.53 / 42.74	28.83 / 29.05	26.68 / 26.85	28.17 / 28.18	25.31 / 25.32
Ours	**68.32 / 69.49**	**66.54 / 66.92**	**31.50 / 31.64**	**30.29 / 30.56**	**29.41 / 29.44**	**25.35 / 25.36**

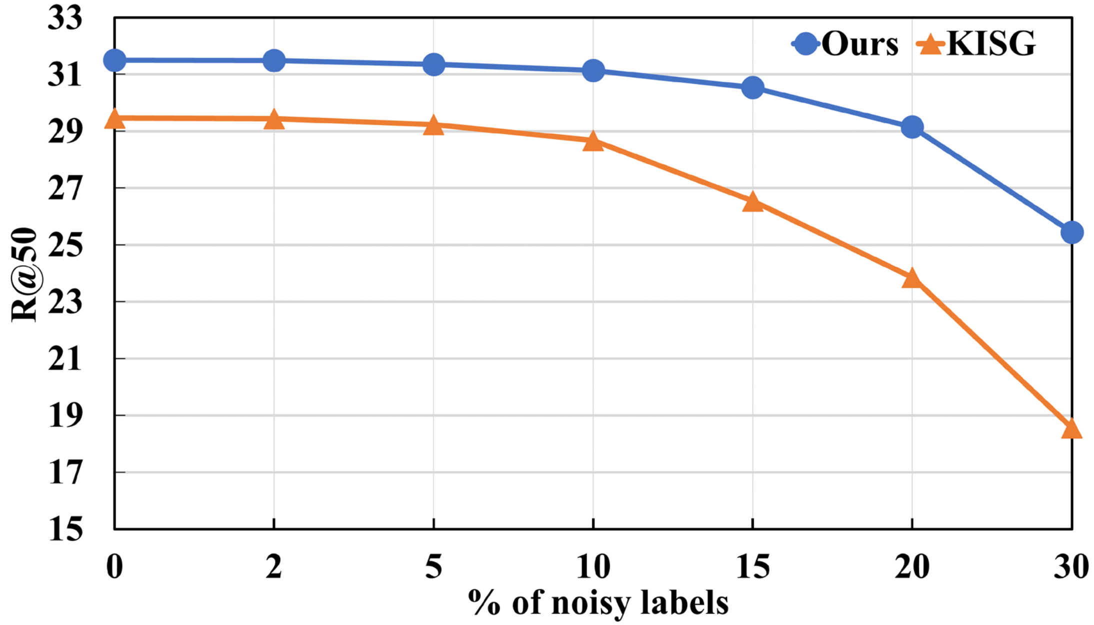
Table 1. Comparison with state-of-the-art 2D scene graph prediction methods re-implemented to work on 3DSSG dataset.



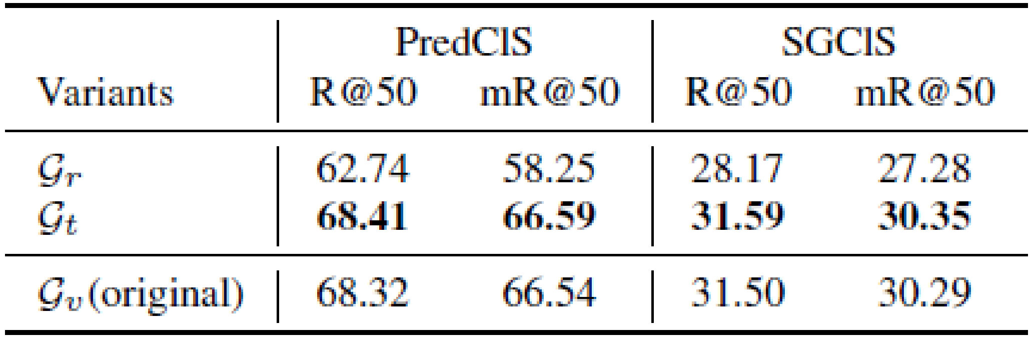
Methods	PredCls		SGCls		SGDet	
	R@50/100	mR@50/100	R@50/100	mR@50/100	R@50/100	mR@50/100
SGPN [37]	57.71 / 58.05	38.12 / 38.67	28.39 / 28.74	22.23 / 22.57	- / -	- / -
EdgeGCN [46]	58.42 / 59.11	38.84 / 39.35	28.58 / 28.93	22.67 / 23.33	- / -	- / -
KISG [47]	64.47 / 64.93	63.19 / 63.52	29.46 / 29.65	28.20 / 28.64	- / -	- / -
Ours	**68.32 / 69.49**	**66.54 / 66.92**	**31.50 / 31.64**	**30.29 / 30.56**	**29.41 / 29.44**	**25.35 / 25.36**

Table 2. Comparison with 3D scene graph prediction methods on the 3DSSG dataset.

Further Analysis



Comparison of our model and KISG on the SGCls task when trained with noisy labels.

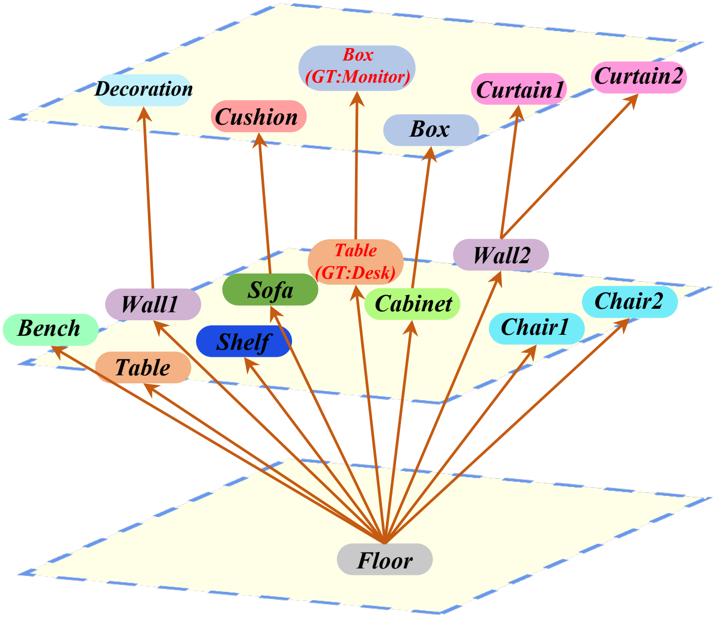


Variants	PredCls		SGCls	
	R@50	mR@50	R@50	mR@50
$\mathcal{G}_r$	62.74	58.25	28.17	27.28
$\mathcal{G}_t$	**68.41**	**66.59**	**31.59**	**30.35**
$\mathcal{G}_v$(original)	68.32	66.54	31.50	30.29

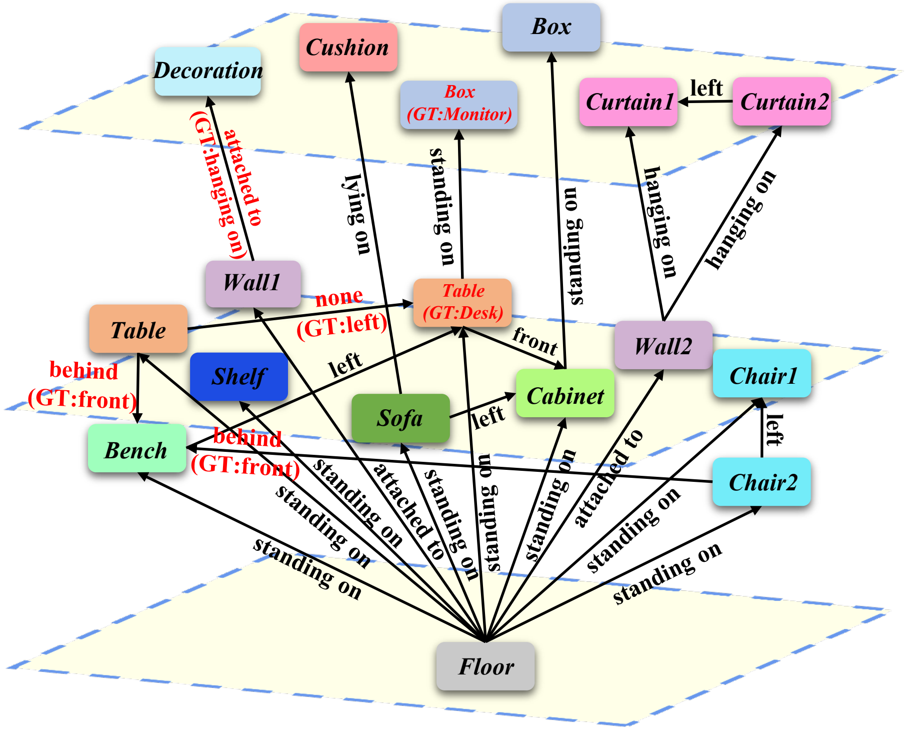
Table 3. Comparison of different variants of the visual graph.

Qualitative Results

Input scene

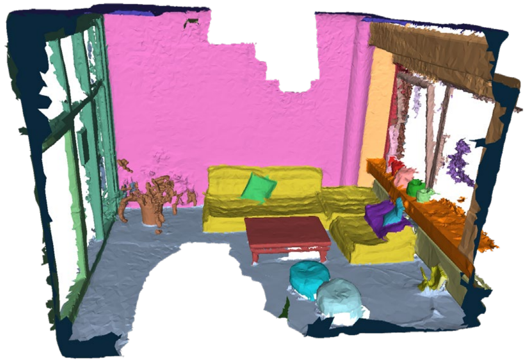


Hierarchical visual graph

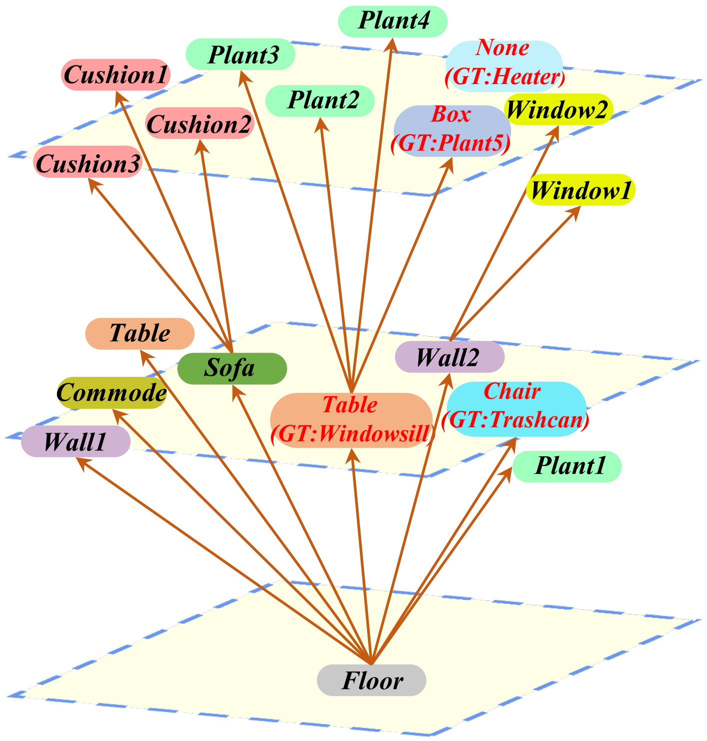


3D scene graph

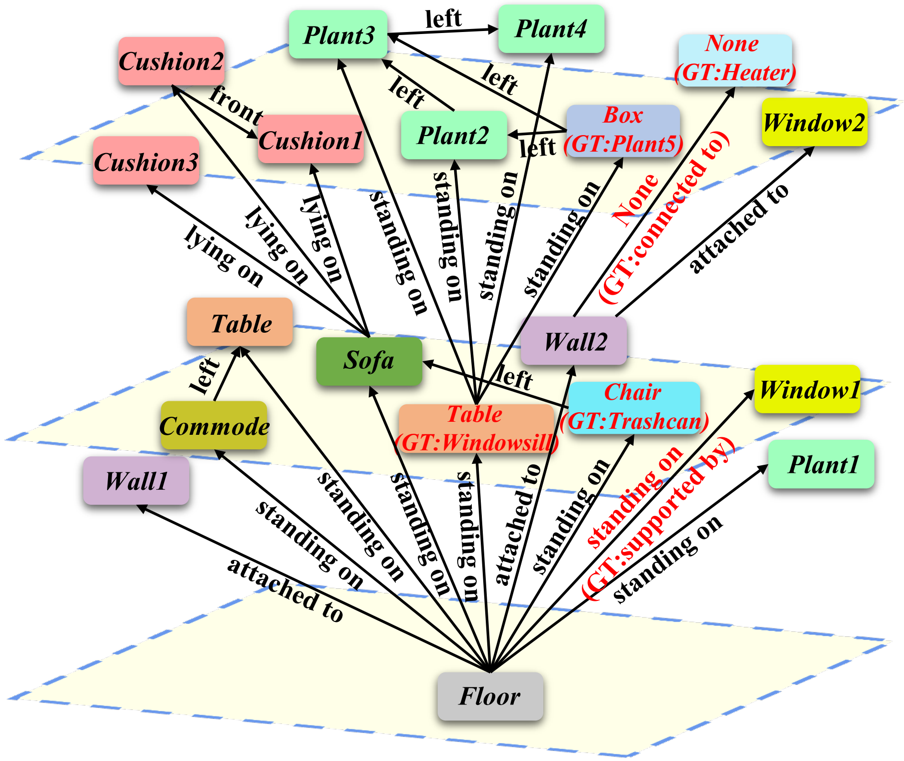
Qualitative Results



Input scene



Hierarchical visual graph



3D scene graph

Conclusion

- Our proposed method explicitly unifies the regular patterns of 3D physical spaces into the deep neural networks to facilitate 3D scene graph prediction.
- We propose a hierarchical symbolic knowledge construction module that exploits extra knowledge as the baseline to admit the hierarchical structure cues of 3D scene.
- We propose a knowledge-guided visual context encoding module to build a hierarchical visual graph and learns the contextualized features by a region-aware graph network.
- A 3D spatial multimodal knowledge accumulation module is proposed to regularize the semantic space of relationship prediction.

Thank you

Please feel free to contact me if you have any questions: mintfeng@hnu.edu.cn