Holistic Pose Graph: Modeling Geometric Structure among Objects in a Scene using Graph Inference for 3D Object Prediction

Jiwei Xiao[1,2], Ruiping Wang[1,2,3], Xilin Chen[1,2]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
[3]Beijing Academy of Artificial Intelligence, Beijing, 100084, China
jiwei.xiao@vipl.ict.ac.cn, {wangruiping, xlchen}@ict.ac.cn

1. Introduction

In this supplementary material, we provide more experiment details mentioned in the main paper. In Section 2, we introduce the implementation details of the settings about learning weights of loss functions. In Section 3, we report the full results of 3D object detection on SUN RGB-D [4] dataset, and the extra experiments for verifying the validity of HPG. In Section 4, we further provide the full results of 3D box estimation. In Section 5, we show more qualitative comparisons. In Section 6, we show more qualitative results about the process of graph inference. In Section 7, we conduct an experiment to explore the complexity of the process of graph inference.

2. Implementation

As mentioned in Loss Functions (Section 3.3) in the main paper, we use a total of five loss functions for training our model. Considering the relative importance and the order of magnitude of each loss function in Equ. 8, we set the $\lambda_{obj} = 1$, $\lambda_{rel} = 1$, $\lambda_{con} = 1$, $\lambda_{co} = 10$, $\lambda_{g} = 100$. For L_{obj}, L_{rel}, L_{con} in Equ. 5-7, each loss function includes four parts of losses for its predicted four parameters.

Specifically, for L_{obj} in Equ. 5, we let $\lambda_{x}^{reg} = 1$, $\forall x \in \{\delta, s\}$, $\lambda_{y}^{reg} = 1$, $\lambda_{y}^{cls} = 1$, $\forall y \in \{\phi, d\}$; for L_{rel} in Equ. 6, we let $\lambda_{\delta_{ij}}^{reg} = 0.3$, $\lambda_{s_{ij}}^{reg} = 0.01$, $\lambda_{\phi_{ij}}^{reg} = 0.3$, $\lambda_{\phi_{ij}}^{cls} = 0.5$, $\lambda_{d_{ij}}^{reg} = 1$, $\lambda_{d_{ij}}^{cls} = 1$; for L_{con} in Equ. 7, we let $\lambda_{x}^{reg} = 1$, $\forall x \in \{\delta_{ij}^{*}, s_{ij}^{*}\}$, $\lambda_{y}^{reg} = 1$, $\lambda_{y}^{cls} = 1$, $\forall y \in \{\phi_{ij}^{*}, d_{ij}^{*}\}$.

3. 3D Object Detection

We report the full results of 3D object detection in Table 3 corresponding to Table 1 in the main paper, which uses the same train/test split and the object labels provided in NYU-37 [3] for fair comparison. The results of [1, 2] are cited from [2]. Please note that the mAP of 4 categories (floor mat, wall, floor and ceiling) is unavailable, because no instance of these categories shows in the test set.

We further conduct an experiment by exploiting HPG in the competitive method CooP [1] for retraining. As shown in Table 1, the performance of (CooP + HPG) gets comprehensive improvements especially on our proposed **HPE** metrics (the last four columns) detailed in Sec.5.4 of the main paper, which further justifies the generalizability and benefits of HPG exploring message passing among objects.

4. 3D Box Estimation

We report the full results of 3D box estimation in Table 4 corresponding to Table 4 in the main paper. The results of [1, 2] are our reproduction, which are under the same experimental settings for fair comparison. Here also note that, similar to Table 3, the Acc and mIoU of 4 categories (floor mat, wall, floor and ceiling) are unavailable.

5. Qualitative Comparisons

We show more qualitative comparisons about 3D object detection task in Figure 1 corresponding to Figure 6 in the main paper. Each group of the qualitative results contains three columns which are the prediction of our baseline (Total3D [2]), our proposed method and ground truth respectively.

6. Qualitative Results of Graph Inference

We show more qualitative results about the process of graph inference in Figure 2 corresponding to Figure 7 in the main paper. To visualize the process of dynamically updating the object pose vividly, we provide 20 typical examples in the form of *".gif"* format files, which are collected in the attached folder named *"graph inference"*.

Table 1. The validity of HPG in CooP [1] on SUN RGB-D dataset.

Method	mAP	mIoU	Acc	$RelAcc$	$PhrAcc$	$RelAcc_I$	$PhrAcc_I$
CooP	21.77	13.70	37.77	20.56	2.82	39.33	4.31
CooP+HPG	24.12	15.35	39.63	27.25	4.29	46.45	7.82

Table 2. Complexity analysis of Graph Inference.

Edges sampling rate	mAP	mIoU	Acc	$RelAcc$	$PhrAcc$	$RelAcc_I$	$PhrAcc_I$
Random 30%	28.72	18.21	49.83	34.22	6.63	55.68	12.80
Random 50%	30.26	19.05	51.46	37.15	8.00	58.78	16.23
Random 70%	31.58	19.83	52.96	38.80	8.66	60.15	18.19
Random 90%	32.26	19.95	53.72	39.85	9.03	60.87	18.73
Fully connected	32.75	20.04	54.27	40.09	9.19	60.83	18.49

7. Complexity analysis of Graph Inference

We explore the impact of different sample rates of the graph edges by random sampling. To model holistic geometric structure, we propose to build HPG as a fully connected graph. Results in Table 2 indicate that **more edges, more constraints, lead to higher performance**. Besides, we carefully calculate the time cost on SUN RGB-D [4] using a GPU of TITAN RTX. Inference time costs: baseline Total3D (**0.1152s/image**), ours (**0.1208s/image**). Benefited from GRU's lightweight and high efficiency, our method does not incur obvious time cost increase.

References

[1] Siyuan Huang, Siyuan Qi, Yinxue Xiao, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. Cooperative holistic scene understanding: Unifying 3d object, layout, and camera pose estimation. *Advances in Neural Information Processing Systems*, pages 206–217, 2018. 1, 2, 3

[2] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 3, 4

[3] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760, 2012. 1

[4] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 1, 2

Table 3. Comparisons of 3D object detection on SUN RGB-D dataset.

Method	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter
CooP [1]	10.47	57.71	15.21	36.67	31.16	0.14	0.00	3.81	0.00	27.67
Total3D [2]	14.51	60.65	17.55	44.90	36.48	0.69	**0.62**	4.93	0.37	32.08
Ours (w/o. HPG)	14.67	60.52	21.28	52.54	35.71	0.47	0.00	9.58	0.02	34.12
Ours	**17.73**	**67.07**	**30.55**	**56.63**	**44.51**	**1.46**	0.00	**13.32**	**0.46**	**34.61**
Method	blinds	desk	shelves	curtain	dresser	pillow	mirror	floor mat	clothes	books
CooP [1]	2.27	19.90	2.96	1.35	15.98	2.53	**0.47**	-	0.00	**3.19**
Total3D [2]	0.00	27.93	3.70	3.04	21.19	**4.46**	0.29	-	0.00	2.02
Ours (w/o. HPG)	0.68	31.90	4.52	0.50	23.00	4.08	0.00	-	0.00	1.52
Ours	**3.81**	**37.82**	**5.00**	**5.48**	**23.40**	4.39	0.04	-	0.00	1.30
Method	fridge	tv	paper	towel	shower curtain	box	whiteboard	person	nightstand	toilet
CooP [1]	21.50	5.20	0.20	**2.14**	**20.00**	2.59	0.16	20.96	11.36	42.53
Total3D [2]	24.42	**5.60**	**0.97**	2.07	**20.00**	2.46	0.61	**31.29**	**17.01**	44.24
Ours (w/o. HPG)	21.36	2.38	0.07	1.25	0.00	1.65	0.20	26.41	12.22	**63.75**
Ours	**25.02**	4.32	0.15	1.64	0.00	**3.15**	**1.42**	23.68	16.93	60.97
Method	sink	lamp	bathtub	bag	wall	floor	ceiling			
CooP [1]	15.95	3.28	24.71	1.53	-	-	-			
Total3D [2]	18.50	5.04	21.15	**2.47**	-	-	-			
Ours (w/o. HPG)	16.86	5.24	**26.51**	1.57	-	-	-			
Ours	**25.70**	**7.15**	17.19	1.17	-	-	-			

Table 4. Comparisons of 3D box estimation on SUN RGB-D dataset. For each column, the left and right results denote the Acc and mIoU individually.

Method	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter
CooP [1]	25.95/9.3	74.28/28.3	33.87/13.2	57.59/19.9	50.95/18.0	5.53/3.0	3.57/2.2	13.29/6.0	2.78/1.1	48.48/16.0
Total3D [2]	36.90/12.6	80.33/30.5	43.74/17.4	68.54/24.9	61.93/22.8	10.05/3.8	3.57/3.0	21.26/8.7	3.47/1.8	59.39/19.8
Ours (w/o. HPG)	37.62/12.8	78.86/30.8	44.26/17.2	74.02/26.8	60.80/21.5	8.54/3.7	**7.14/3.2**	27.91/10.6	2.08/1.6	60.61/20.4
Ours	**43.57/14.7**	**81.62/33.0**	**50.70/20.5**	**77.62/29.3**	**67.90/24.0**	**10.55/4.2**	**7.14**/2.9	**42.19/14.2**	**6.94/2.4**	**65.45/20.5**
Method	blinds	desk	shelves	curtain	dresser	pillow	mirror	floor mat	clothes	books
CooP [1]	**4.55**/2.4	42.28/14.4	16.11/6.2	4.17/4.0	29.13/9.6	10.28/3.8	2.67/1.0	-	0.00/0.00	3.23/2.0
Total3D [2]	0.00/1.6	53.93/19.2	23.77/8.8	10.42/4.6	44.34/15.9	17.21/6.7	2.67/2.1	-	0.00/0.00	**9.68/3.5**
Ours (w/o. HPG)	**4.55**/2.8	55.62/19.6	24.36/8.3	12.50/6.1	45.63/**17.0**	18.01/6.8	1.33/1.0	-	0.00/10.7	7.53/2.7
Ours	**4.55/3.1**	**60.29/21.6**	**30.06/10.8**	**18.06/6.9**	**48.54**/16.4	**23.09/8.6**	**5.33/2.9**	-	**1.00/16.1**	8.60/3.2
Method	fridge	tv	paper	towel	shower curtain	box	whiteboard	person	nightstand	toilet
CooP [1]	39.62/14.7	15.25/5.3	1.42/0.6	2.35/0.9	0.00/7.9	7.09/2.6	2.68/2.0	26.19/10.3	20.95/7.9	58.00/22.4
Total3D [2]	46.54/17.4	**24.58**/8.1	6.38/2.0	10.59/4.3	**1.00**/38.5	16.45/6.3	**10.07**/3.7	**61.90/22.3**	39.13/**14.6**	68.00/25.9
Ours (w/o. HPG)	46.54/18.7	16.95/5.7	2.84/1.3	16.47/5.8	**1.00/44.9**	14.61/5.4	5.37/3.2	42.86/15.6	34.78/12.7	68.67/**28.9**
Ours	**49.06/19.0**	21.19/**8.5**	**6.94/2.2**	**21.18/6.4**	**1.00**/37.4	**17.16/6.6**	**10.07/4.3**	47.62/17.8	**40.32**/13.7	**74.67**/28.3
Method	sink	lamp	bathtub	bag	wall	floor	ceiling			
CooP [1]	24.36/9.7	17.84/6.4	37.25/11.9	7.69/2.2	-	-	-			
Total3D [2]	45.89/17.2	19.46/7.6	**47.06**/17.3	**23.08/7.9**	-	-	-			
Ours (w/o. HPG)	39.38/15.0	21.35/7.7	45.10/**17.7**	19.23/7.4	-	-	-			
Ours	**49.00/18.2**	**23.78/9.0**	31.37/13.1	**23.08**/7.1	-	-	-			

Figure 1. More qualitative comparisons with [2] about 3D object detection task on SUN RGB-D dataset.

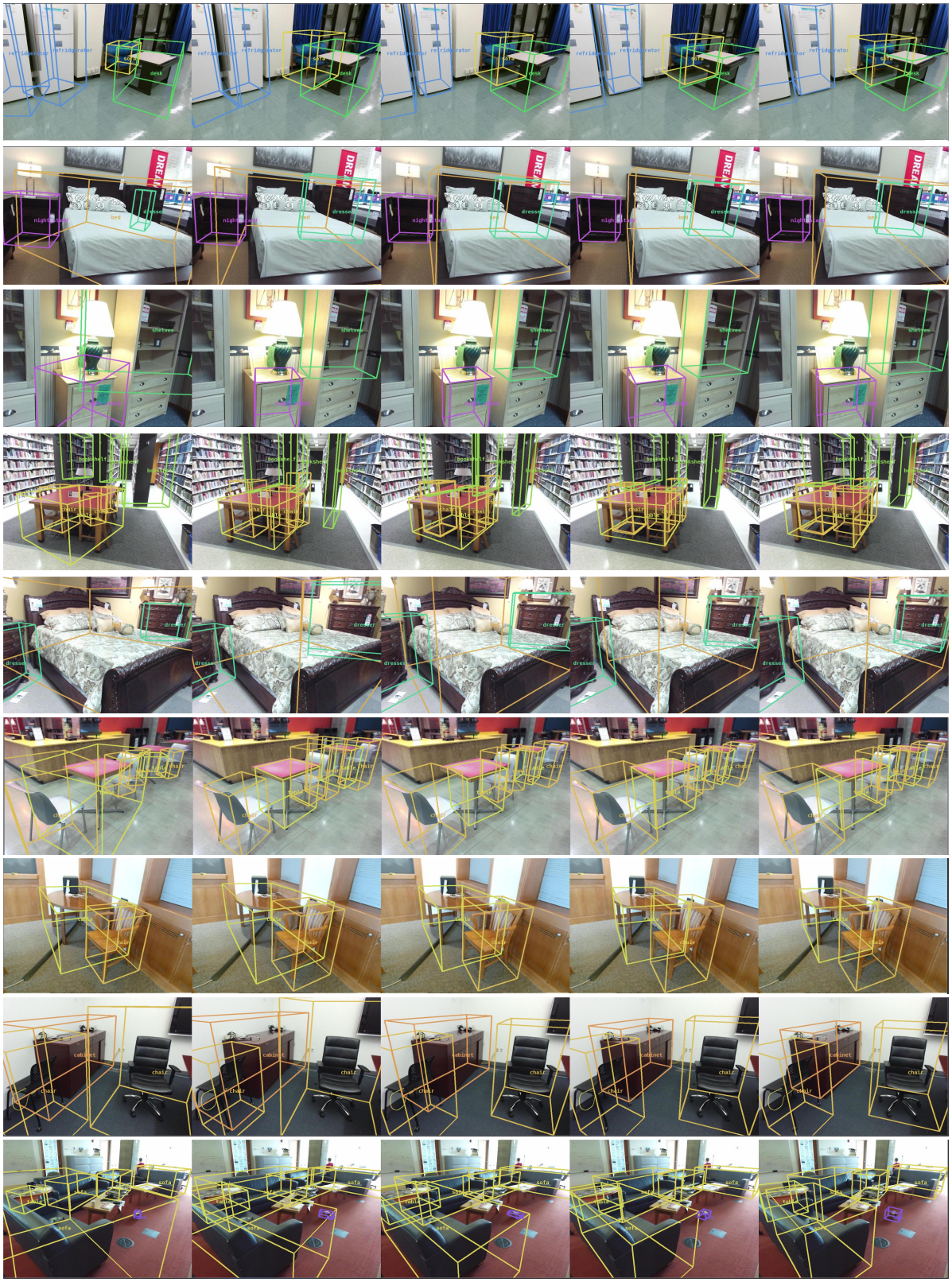

Figure 2. Visualization of the intermediate results during graph inference. T denotes the iterations of the message passing process.