Back to Reality: Weakly-supervised 3D Object Detection with Shape-guided Label Enhancement

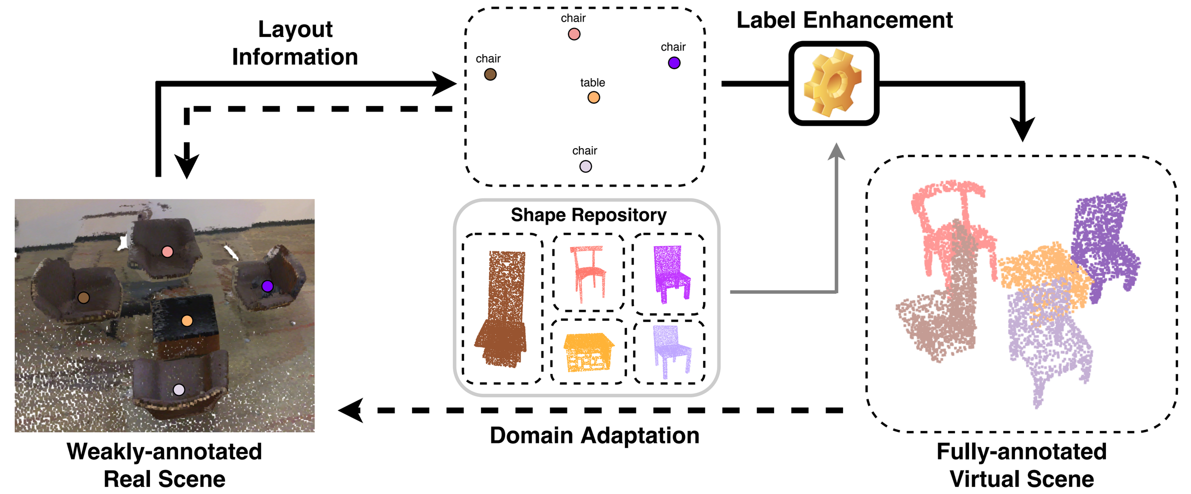
Xiuwei Xu[1,2], Yifan Wang[2], Yu Zheng[1,2], Yongming Rao[1,2], Jie Zhou[1,2], Jiwen Lu[1,2]
[1] Department of Automation, Tsinghua University, China
[2] Beijing National Research Center for Information Science and Technology, China

Motivation

- The annotation time for 3D object detection is a huge obstacle preventing its practical application. Compared to 2D counterpart, labeling a bounding box for 3D point cloud is takes much more time (more than 100s per object). Considering time-accuracy tradeoff, position-level weakly-supervised method is a promising topic.

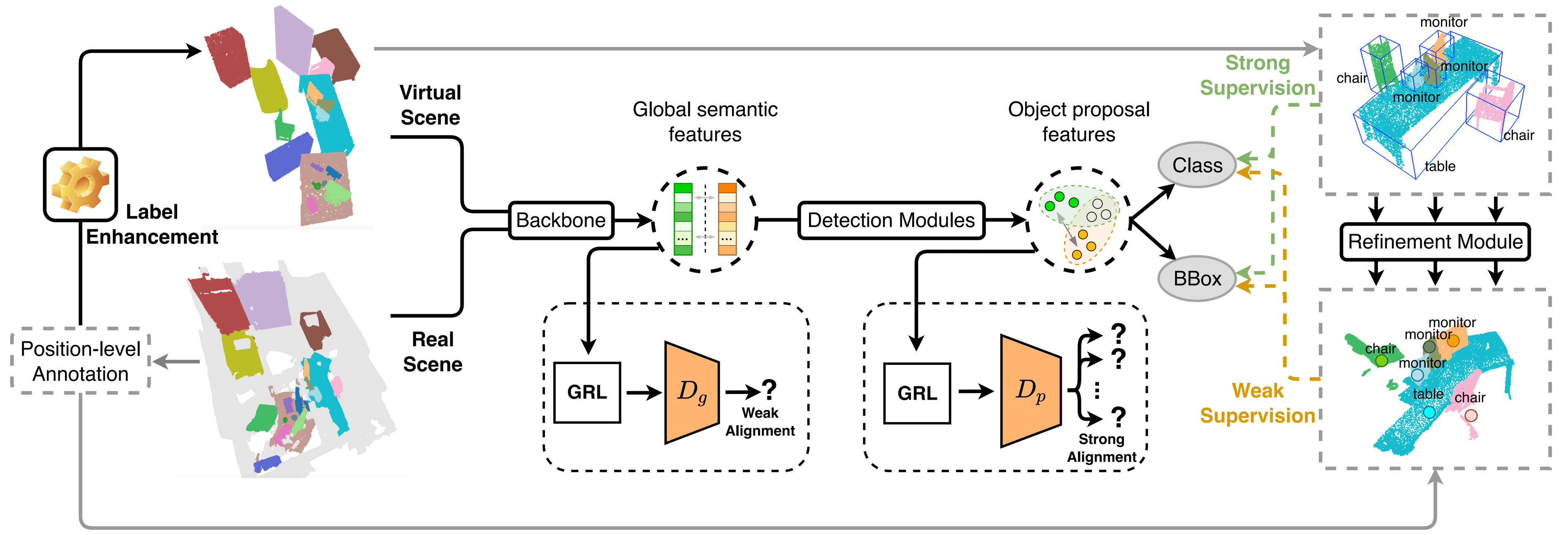
Annotation	BBox [22]	S-L [35]	P-L [23]	P-L(BR)
Time(s per object)	110	1	5	5
mAP@0.25(%)	54.2	<20	32.4	47.0

- To fully exploit the information contained in the position-level annotations, we consider them as the coarse layout of scenes and convert them into virtual scenes with the guide of synthetic 3D shapes for better supervision on object detection task.

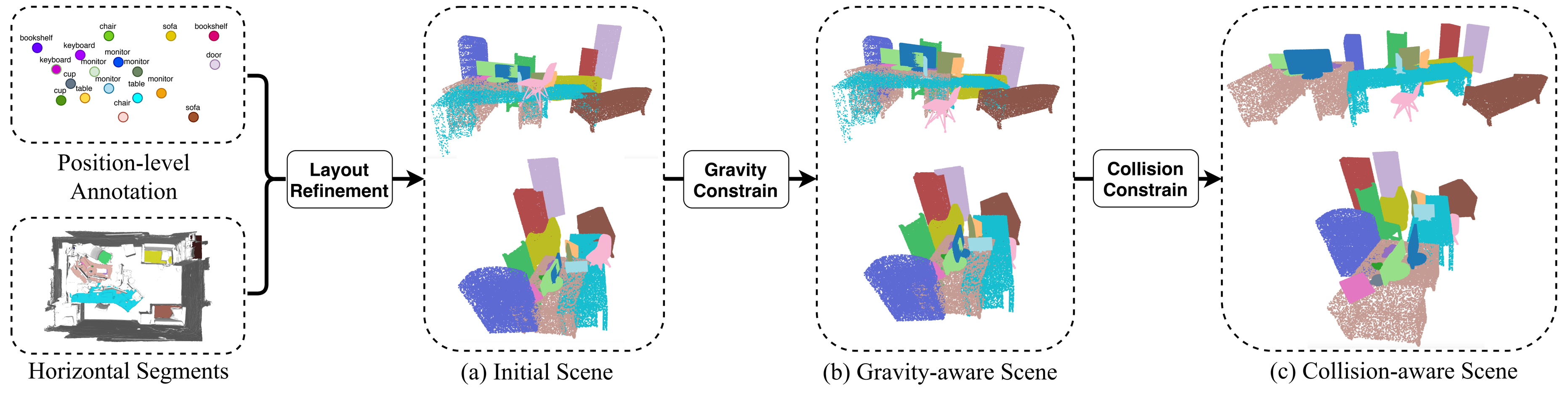


Approach

- The framework of BackToReality:



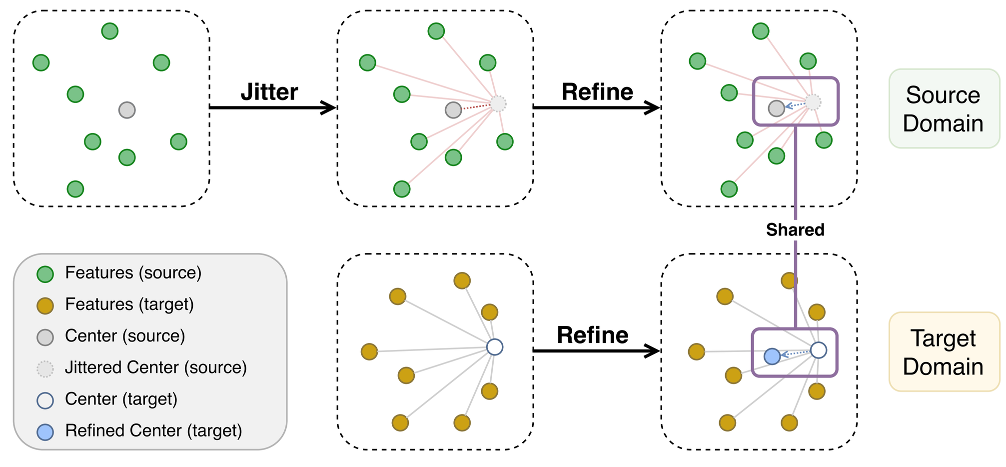
- **Shape-guided Label Enhancement:** we assemble synthetic 3D shapes according to the layout information provided by the weak annotations, and apply physical constraints on the constructed virtual scenes to remedy the information loss from boxes to centers.



- **Virtual-to-Real Domain Adaptation:** given real scenes with position-level annotations and virtual scenes with box annotations, we model the learning of 3D object detector as a domain adaptation problem.

$$\max_{D} \min_{B} J = L_{sup}(B) - L_{adv}(B, D)$$

- The position-level annotations are not precise due to labeling error. Thus we make use of the perfect virtual labels to refine them.



We jitter the centers in virtual scenes to imitate the labeling error in real scenes.

Then we train a PointNet-like module to predict the offset of labeled centers.

$$p(c) = \mathrm{MLP}_2\left\{\max_{i \in N(c)} \{\mathrm{MLP}_1[f_i; c_i - c]\}\right\}$$

- As virtual scenes generated from weak labels lack background or fine-grained layout, the domain gap between real and virtual scenes is very large. We solve this by strong-weak feature alignment.

$$L_{adv}(O, D) = L_{global} + L_{proposal}$$

$$L_{global} = -\sum_{i=1}^{B} (1-p_i)^{\gamma} log(p_i),\ \gamma > 1 \qquad L_{proposal} = \sum_{i=1}^{B}\sum_{j=1}^{N} s_{ij}(1-p_{ij})^2$$

Experiments

- Visualization for virtual scenes:

(a) Real Scenes (b) Cared Objects (c) Mesh-version Virtual Scenes (d) Point-version Virtual Scenes

- 3D object detection results:

	Setting	batht.	bed	bench	bsf.	bot.	chair	cup	curt.	desk	door	dres.	keyb.	lamp	lapt.	monit.	n.s.	plant	sofa	stool	table	toil.	ward.	mAP@0.25
VoteNet	FSB [27]	66.8	86.2	24.4	**55.6**	0.0	88.3	0.0	48.5	62.8	45.8	24.1	0.1	47.2	5.2	62.1	73.2	13.4	88.7	35.1	62.6	94.6	7.8	45.1
	WSB	21.9	46.9	0.3	2.3	0.0	53.7	0.0	0.9	32.1	1.0	6.6	0.1	0.2	0.1	1.8	53.6	0.1	57.0	4.6	6.4	19.7	0.0	14.1
	WS3D† [23]	22.0	58.5	10.3	5.8	0.0	60.4	0.0	4.1	26.7	3.2	1.6	0.0	14.0	0.6	18.6	46.3	0.4	32.7	11.8	23.5	65.0	0.0	18.4
	WSBP_P	43.2	58.0	2.4	16.1	0.0	75.1	0.7	7.9	54.2	6.4	7.1	2.3	35.2	18.4	12.8	64.0	4.4	68.5	20.2	22.0	71.6	5.2	27.1
	WSBP_M	45.0	49.6	5.5	18.5	0.0	62.7	2.9	11.4	49.6	6.9	2.5	1.0	30.0	7.6	21.4	64.8	7.3	79.6	23.1	35.2	80.9	2.2	27.6
	BR_P(Ours)	51.2	73.0	16.4	27.1	0.1	70.3	0.0	8.3	44.5	7.3	16.0	1.5	40.2	7.7	42.1	50.8	7.4	67.1	10.7	39.0	88.4	18.1	31.2
	BR_M(Ours)	57.1	80.4	14.3	31.7	0.0	77.4	0.0	13.2	49.7	11.3	14.8	1.0	43.5	6.0	56.5	65.0	10.6	80.2	26.9	44.2	91.4	6.5	35.5
GroupFree3D	FSB [22]	**86.2**	87.5	16.3	49.6	0.6	**92.5**	0.0	**70.9**	**78.5**	**53.5**	**56.0**	6.4	**68.2**	11.5	**81.5**	**88.5**	15.2	88.2	**45.6**	**65.0**	**99.7**	**31.2**	**54.2**
	WSB	75.0	75.7	4.3	17.2	0.0	81.4	0.0	3.5	34.0	4.7	3.2	2.1	46.6	3.3	45.8	52.8	8.3	71.0	15.7	18.1	90.8	0.7	29.7
	WS3D† [23]	71.9	78.3	0.9	20.2	0.8	79.2	1.0	2.9	47.6	7.7	10.6	19.2	41.6	13.5	65.6	41.2	0.8	74.6	17.7	26.3	88.9	1.7	32.4
	WSBP_P	71.9	77.1	7.7	25.2	**3.0**	80.6	0.4	3.2	50.1	10.5	36.3	17.0	52.9	**30.3**	59.9	63.8	9.6	78.2	28.4	25.3	93.3	14.4	38.2
	WSBP_M	81.8	82.6	0.0	35.0	0.0	77.5	0.4	27.1	38.4	7.6	22.3	9.7	44.3	24.4	65.4	76.5	5.5	62.4	34.7	28.7	**99.7**	5.4	37.7
	BR_P(Ours)	72.3	73.5	**45.8**	27.7	0.0	77.2	**8.2**	30.8	35.0	17.8	51.7	0.3	64.2	25.0	63.5	66.6	23.8	86.7	33.9	37.6	98.3	5.2	43.0
	BR_M(Ours)	85.3	**90.9**	8.8	34.3	1.9	80.0	7.7	24.7	58.0	20.8	45.4	**31.3**	64.4	25.8	67.5	76.7	**27.3**	**91.4**	43.3	46.7	94.8	8.3	47.1

- Ablation Study:

Gravity Constrain	Collision Constrain	Density Control	mAP@0.25
			26.3
✓			27.2
✓	✓		28.5
✓	✓	✓	**31.2**

Global Alignment	Proposal Alignment	Center Refinement	mAP@0.25
			24.2
✓			28.7
	✓		27.4
✓	✓		30.2
✓	✓	✓	**31.2**

Method	Error Rate 10%	20%	30%	40%	50%
WSB	29.7	26.8	25.0	22.3	19.7
BR_M(Ours)	**47.1**	**46.0**	**43.9**	**43.1**	**41.2**