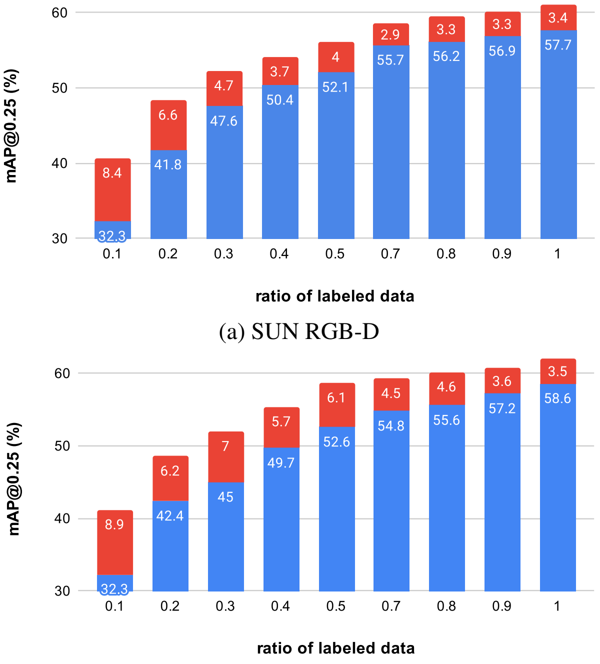
SESS: Self-Ensembling Semi-Supervised 3D Object Detection—Supplementary Material

In this appendix, we provide performance comparison between SESS and VoteNet with more diverse ratios of labeled data on the SUN RGB-D and ScaNetV2 val sets in Sec. A. We also provide additional evaluation metric (*i.e.* mAP@0.5 IoU) for both inductive and transductive semi-supervised learning in Sec. B. In Sec. C, we report per-class average precision on the SUN RGB-D and ScanNetV2 val set. Finally, more qualitative results are shown in Section D.

A. Additional Label ratios



(a) SUN RGB-D

(b) ScanNetV2

Figure 1: Comparison to VoteNet with more ratios of labeled data on the SUN RGB-D and ScanNetV2 val sets. The blue columns denote the performances of VoteNet, and the red columns denote the improved performance of SESS over VoteNet.

More ratios (*i.e.* 80% and 90%) of labeled data are included in the performance comparison of our SESS to the fully-supervised VoteNet on two datasets. The comparison results are illustrated in Figure 1. As can be seen from the figures, the performance margin (compared to the performance of using 100% labeled data) becomes smaller when the ratio of labeled data increases. This is because the same type of scenes (*e.g.* classrooms) share similar layout and/or objects, and thus the contribution of new labeled data to model training might be minor when similar types of samples/scenes have been seen by the model.

Table 1: Inductive leaning on SUN RGB-D and ScanNetV2 val sets compared with the fully supervised VoteNet, evaluated by mAP@0.5 IoU. The percentage indicates the ratio of labeled data for training.

Dataset	Model	10%	20%	30%	40%	50%	70%	100%
SUNRGB-D	VoteNet	10.6	14.7	23.3	25.6	27.2	30.0	31.1
	SESS	**14.4**	**20.6**	**28.5**	**29.0**	**30.6**	**33.4**	**37.3**
ScanNetV2	VoteNet	11.9	21.2	22.5	27.7	28.9	30.9	33.5
	SESS	**18.6**	**26.9**	**27.4**	**31.5**	**34.2**	**35.5**	**38.8**

Table 2: Transductive leaning on SUN RGB-D and ScanNetV2 unlabeled training sets compared with the fully supervised VoteNet, evaluated by mAP@0.5 IoU. The percentage indicates the ratio of labeled data for training.

Dataset	Model	10%	20%	30%	40%	50%	70%
SUNRGB-D	VoteNet	10.3	15.3	23.4	25.5	25.0	29.9
	SESS	**15.8**	**20.1**	**27.4**	**27.2**	**29.2**	**36.7**
ScanNetV2	VoteNet	13.8	25.3	28.6	32.7	35.2	38.3
	SESS	**23.2**	**31.3**	**34.3**	**37.6**	**41.6**	**42.6**

B. Additional Evaluation Metric

We additionally evaluate mean average precision with an IoU threshold of 0.5 on the SUN RGB-D and ScanNetV2 for both inductive (see Table 1) and transductive (see Table 2) semi-supervised 3D object detection. Consistent with the evaluation at an IoU threshold of 0.25, our SESS significantly outperforms the fully supervised VoteNet under different ratios of labeled data for both inductive and transductive learning.

Table 3: Per-class mAP@0.25 IoU on SUN RGB-D val set, with 100% training samples. The upper table lists the results obtained by five fully-supervised methods, and the lower table lists the results of our proposed semi-supervised method.

Method	bathtub	bed	bookshelf	chair	desk	dresser	nightstand	sofa	table	toilet	mAP
DSS	44.2	78.8	11.9	61.2	20.5	6.4	15.4	53.5	50.3	78.9	42.1
COG	58.3	63.7	31.8	62.2	**45.2**	15.5	27.4	51.0	51.3	70.1	47.6
2D-driven	43.5	64.5	31.4	48.3	27.9	25.9	41.9	50.4	37.0	80.4	45.1
F-PointNet	43.3	81.1	33.3	64.2	24.7	**32.0**	58.1	61.1	51.1	90.9	54.0
VoteNet	74.4	83.0	28.8	75.3	22.0	29.8	62.2	64.0	47.3	90.1	57.7
SESS	**76.9**	**84.8**	**35.4**	**75.8**	29.3	31.3	**66.9**	**66.4**	**51.8**	**92.3**	**61.1**

Table 4: Per-class mAP@0.25 IoU on ScanNetV2 val set, with 100% training samples. The upper table lists the results from two fully-supervised methods, and the lower table lists the results of our proposed semi-supervised method.

Method	cabin.	bed	chair	sofa	table	door	wind.	bkshf	pic.	cntr	desk	curt.	fridg.	showr.	toilet	sink	bath	ofurn.	mAP
3DSIS	19.8	69.7	66.2	71.8	36.1	30.6	10.9	27.3	0.0	10.0	46.9	14.1	53.8	36.0	87.6	43.0	84.3	16.2	40.2
VoteNet	36.3	87.9	**88.7**	89.6	58.8	47.3	38.1	44.6	7.8	**56.1**	71.7	**47.2**	45.4	57.1	94.9	**54.7**	92.1	37.2	58.6
SESS	**41.1**	**88.1**	85.9	**91.7**	**64.5**	**52.1**	**40.4**	**51.4**	**11.8**	51.9	**74.9**	45.9	**59.6**	**73.3**	**98.3**	53.9	**93.0**	**39.5**	**62.1**

C. Per-class Evaluation

We respectively report per-class average precision on 10 classes of SUN RGB-D and 18 classes of ScanNetV2 in Table 3 and 4, using all the training samples. Our SESS is superior than the fully supervised VoteNet on each class of SUN RGB-D and 14 classes of ScanNetV2 with the assistance of the proposed pertubation scheme and consistency losses.

D. More Qualitative Results and Discussions

Figure 2 and 3 demonstrate additional qualitative results on the SUN RGB-D and ScanNetV2 val datasets, respectively. As can be seen from the four examples in Figure 2, the heavy occlusion (*e.g*. the chairs at the back rows in the classroom), partial visibility (*e.g*. the leftmost cabinet in the bedroom), and extreme sparsity (*e.g*. the rightmost chair in the study space) make the detection on SUN RGB-D very difficult. Some of them are even hard for human to recognize without the reference of the associated RGB images, such as the leftmost chair in the second row in the classroom and the rightmost chair in the study space. Both VoteNet and our SESS fail to detect these extremely challenging objects that come with no or few representative points. However, it is interesting to see that our SESS successfully detect most of the objects in these challenging scenarios, including those unannotated objects such as the chairs in the back of the classroom, and the table in front of the bed in the bedroom.

In Figure 3, we also show four more examples covering various scenarios on ScanNetV2 dataset. Objects with strong geometric cues (*e.g*. table, chair, bed, desk *etc*.) are easy to detect since both strongly supervised VoteNet and our SESS rely on only the geometric data (*i.e*. XYZ coordinates). In contrast, objects without explicit geometric features (*e.g*. door, picture, window) are difficult to recognize. Despite the challenge, our SESS is able to detect most of the difficult objects, such as bookshelves in the library and doors in the lounge. We argue that the proposed consistency losses, which encode not only geometric but also semantic information, guide the model to achieve better localization of the 3D bounding boxes.

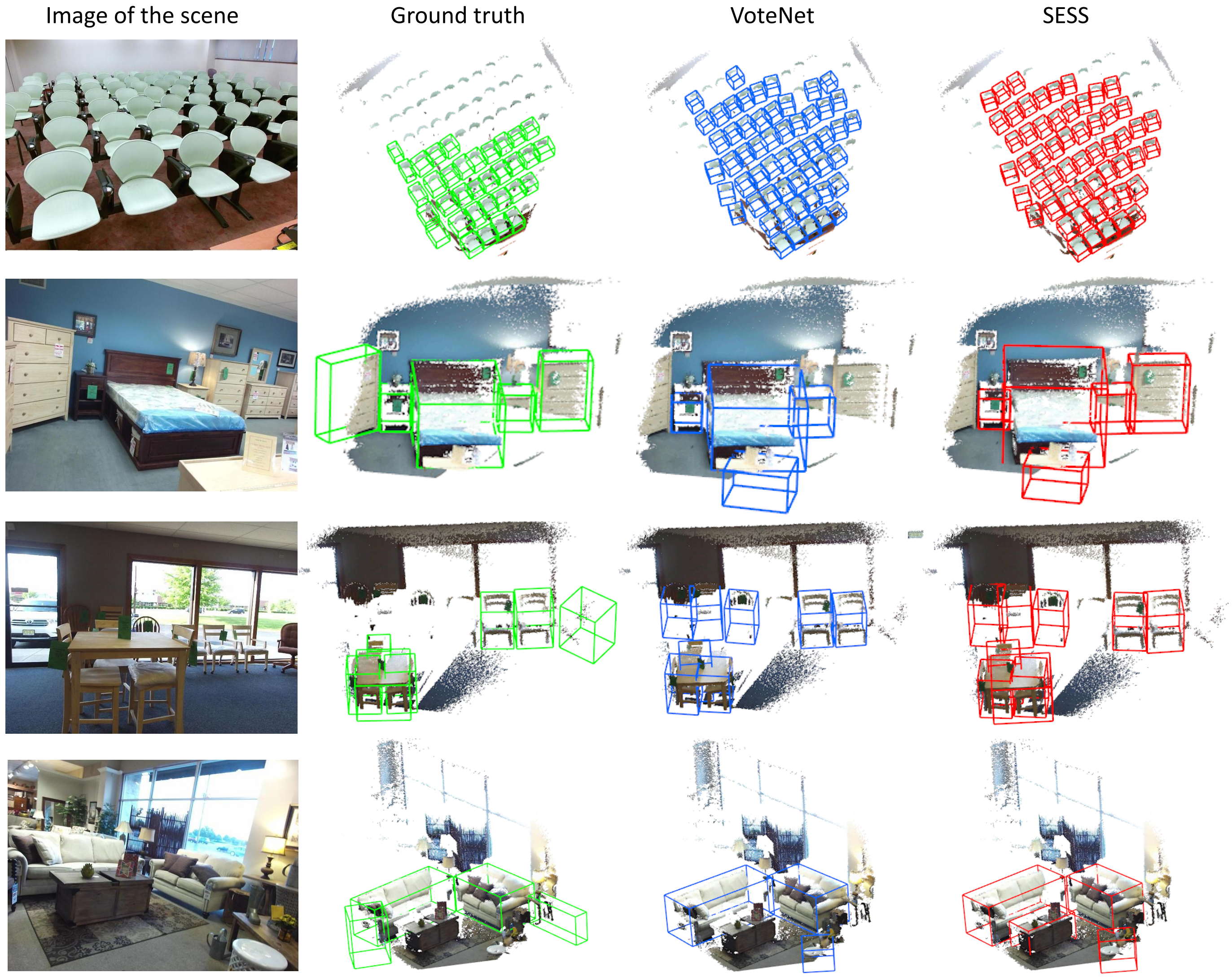


Figure 2: Additional Qualitative comparison between the fully-supervised VoteNet and the proposed SESS on SUN RGB-D *val* set, using 100% training samples. Four scene types are illustrated from the upper to bottom, they are *classroom*, *bedroom*, *study space*, and *living room*.

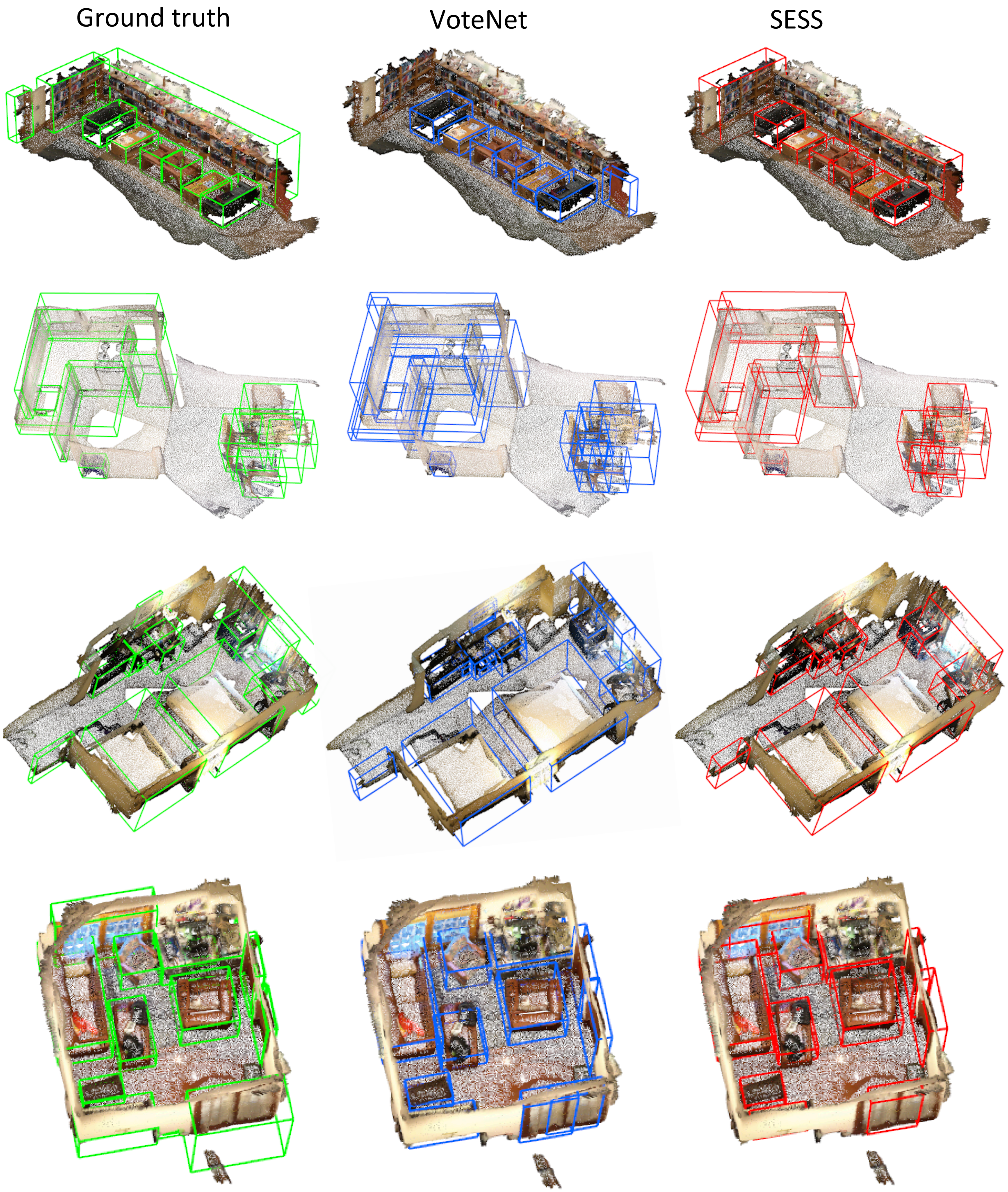


Figure 3: Additional Qualitative comparison between the fully-supervised VoteNet and the proposed SESS on ScanNetV2 *val* set, using 100% training samples. Four scene types are illustrated from the upper to bottom, they are *library*, *kitchen*, *hotel*, and *lounge*.