Supplementary Materials for
HyperDet3D: Learning a Scene-conditioned 3D Object Detector

A. Technical Details

Here we introduce some techinical details on the proposed HyperDet3D, including the processing of obtaining $\boldsymbol{b}$, and some hyper-parameters for the experiments on 2 datasets.

A.1. Processing of obtaining biases

For bias $\boldsymbol{b}$, we maintain $\boldsymbol{Z}^a$ and $\boldsymbol{Z}^s$ which are the same as in obtaining $\boldsymbol{W}$. The scene-specific and scene-agnostic hypernetworks both contain a single linear layer. Before the second attention, we take the average results of scene-specific and scene-agnostic vectors along the feature dimension.

A.2. Hyper-parameters

Following [3], the number of decoder layers for ScanNet [2] and SUN RGB-D [5] are 12 and 6 respectively. We set C_a=256 and n=256 for both datasets. We set C_s=253 and 285 for ScanNet and SUN RGB-D respectively.

B. Other Experimental Results

B.1. Detailed Results on SUN RGB-D

As shown in Table 1, we detail the per-category results on the SUN RGB-D [5] dataset. We observe the obvious improvements on category *bed*, *bookshelf*, *nightstand* and *toilet* which are more conditioned on its corresponding scenes (beddroom, washroom, etc.). The exception is the *bathtub* category with performance drop of -4.9%, which might be attributed to the extremely scarcity of its annotations in the dataset (the least one in the statistical distribution shown in the paper [5]).

B.2. Detailed Results on Cross-dataset Evaluation

For the cross-dataset experiments where we pretrained HyperDet3D on SUN RGB-D and finetuned on ScanNet v2, we display the detailed per-category results in Table 2, including the extra 10 novel categories.

B.3. Ablation Study on Cross-dataset Evaluation

The removal of scene-specific knowledge and MSA→SSA brings -3.0% and -3.4% on mAP_8 respectively. We infer the expressiveness of MSA contributes more than scene-specific knowledge when tackling domain gap.

C. More Visualization Results

We have additionally illustrated qualitative results on ScanNet v2 and SUN RGB-D. The results on ScanNet v2 are shown in Figure 1. The results on SUN RGB-D are shown in Figure 2.

References

[1] Bowen Cheng, Lu Sheng, Shaoshuai Shi, Ming Yang, and Dong Xu. Back-tracing representative points for voting-based 3d object detection in point clouds. In *CVPR*, pages 8963–8972, 2021. 2

[2] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828—-5839, 2017. 1

[3] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *ICCV*, pages 2949–2958, 2021. 1, 2

[4] Charles R. Qi, Or Litany, Kaiming He, and Leonidas J. Guibas. Deep hough voting for 3d object detection in point clouds. In *ICCV*, pages 9277–9286, 2019. 2

[5] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, pages 567–576, 2015. 1, 2

[6] Zaiwei Zhang, Bo Sun, Haitao Yang, and Qixing Huang. H3dnet: 3d object detection using hybrid geometric primitives. In *ECCV*, pages 311–329, 2020. 2

Table 1. 3D object detection results on SUN RGB-D V1 validation dataset. We show per-category results of mean average precision (mAP) with 3D IoU threshold 0.5 as proposed in [5], and mean of AP across all semantic classes with 3D IoU threshold 0.5.

	bathtub	bed	bookshelf	chair	desk	drser	nightstand	sofa	table	toilet	mAP
Votenet [4]	45.4	53.4	6.8	56.5	5.9	12.0	38.6	49.1	21.3	68.5	35.8
H3DNet [6]	47.6	52.9	8.6	60.1	8.4	20.6	45.6	50.4	27.1	69.1	39.0
BRNet [1]	55.5	63.8	9.3	61.6	10.0	**27.3**	53.2	56.7	28.6	70.9	43.7
GF3D [3]	**64.0**	67.1	12.4	**62.6**	**14.5**	21.9	49.8	58.2	**29.2**	72.2	45.2
Ours	59.1	**69.4**	**21.1**	62.1	11.3	14.0	**57.4**	**61.3**	27.6	**89.8**	**47.3**

Table 2. 3D object detection results on the ScanNet V2 validation dataset. We show per-category results of mean average precision (mAP) with 3D IoU threshold 0.5 as proposed in [5], and mean of AP across all semantic classes with 3D IoU threshold 0.5.

	cab	bed	chair	sofa	tabl	door	wind	bkshf	pic	cntr	desk	curt	fridg	showr	toil	sink	bath	ofurn	mAP
GF3D [3]	5.3	66.6	21.3	46.9	17.8	3.3	4.5	0.4	0	8.0	25.6	6.7	22.3	0	54.6	11.7	48.6	0.4	19.1
Ours	10.8	78.9	22.7	58.0	16.0	2.9	2.2	2.4	0.5	13.3	40.1	11.6	7.4	0	58.9	0.5	71.4	2.3	22.2

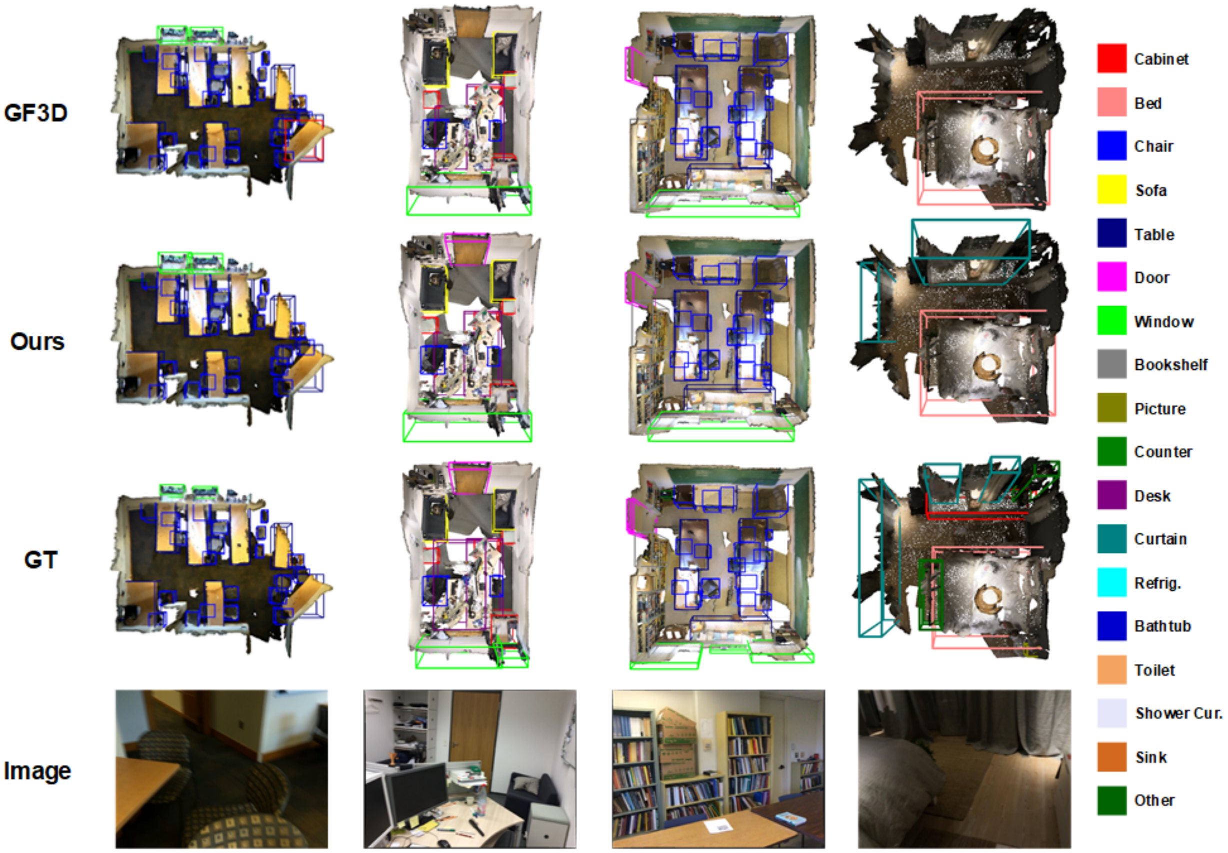

Figure 1. Qualitative detection results on the ScanNet V2 validation set. (*Best viewed in color.*)

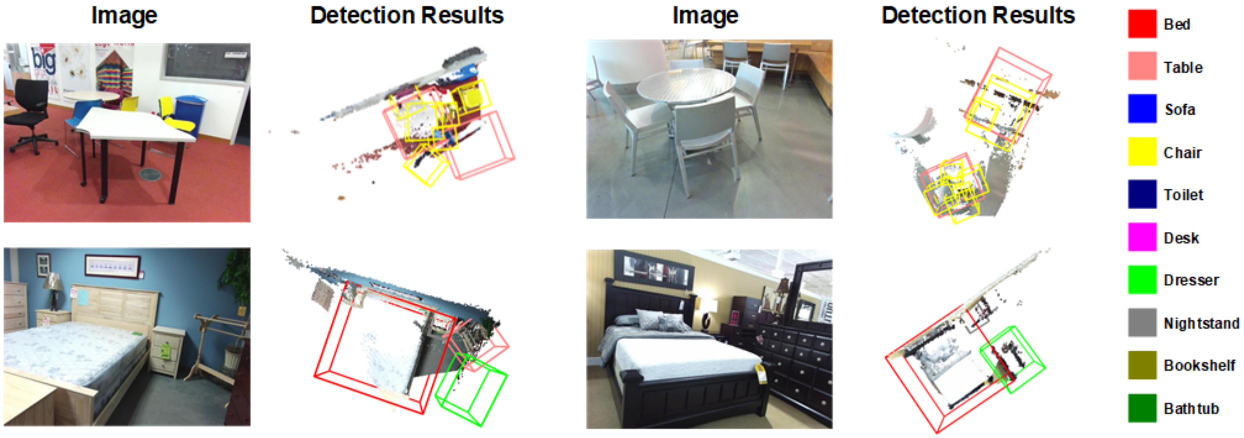

Figure 2. Qualitative detection results on the SUN RGB-D V1 validation set. (*Best viewed in color.*)