DisARM: Displacement Aware Relation Module for 3D Detection

Yao Duan Chenyang Zhu Yuqing Lan Renjiao Yi Xinwang Liu Kai Xu*
National University of Defense Technology

1. Appendix

The Appendix mainly includes subsection **Training Details** and subsection **More Results**. We describe the detailed training strategies of different methods equipped with our DisARM in **Training Details**. What's more, the quantitative and qualitative results on ScanNet V2 and SUN RGB-D datasets are also shown in **More Results**.

A. Training Details

VoteNet+DisARM. Our DisARM plugged into VoteNet [7] is end-to-end optimized using the AdamW [5] optimizer with the batch size of 8 for ScanNet V2 and 16 for SUN RGB-D. The initial learning rate is 0.008 and the network is trained for 220 epochs on both datasets. The Cosine Annealing [4] is adopted as the learning rate schedule. We implement our method on MMDetection3D [2] with one NVIDIA TITAN V GPU.

BRNet+DisARM. BRNet [1] equipped with our DisARM is end-to-end optimized using the AdamW [5] optimizer with the batch size of 8. The initial learning rate is 0.005 and the weight decay is 0.01. The network is trained for 64 epochs on ScanNet V2 dataset. The Cosine Annealing [4] is adopted as the learning rate schedule. We implement our method on MMDetection3D [2] with one NVIDIA TITAN V GPU.

GroupFree3D+DisARM. The GroupFree3D [3] equipped with DisARM also can be trained end-to-end with the AdamW [5] optimizer. The batch size is 8. The initial learning rate is 0.006 and the network is trained for 120 epochs on ScanNet V2 dataset. The decay steps are 56 and 68. GroupFree3D [3] has different backbone settings and object candidate proposals, and thus there are different model sizes. For the lightweight settings with 6-layer decoder, 256 object candidates and 12-layer decoder, 256 object candidates, We train the GroupFree3D+DisARM with one NVIDIA TITAN V GPU. For the more complex settings of 12-layer decoder with 256 candidates and 512 candidates respectively, we train the GroupFree3D+DisARM with one NVIDIA GeForce RTX 3090 GPU. Note that the feature dimension F of candidate proposal is 288, which is different from other backbone methods.

H3DNet+DisARM. We plug DisARM into H3DNet [9], which is trained with the AdamW [5] optimizer and batch size of 3. The initial learning rate is 0.008 and the weight decay is 0.01. The network is trained for 56 epochs on ScanNet V2 dataset. The decay steps are 24 and 32. We implement our method on MMDetection3D [2] with one NVIDIA TITAN V GPU.

imVoteNet+DisARM. imVoteNet [6] equipped with our DisARM is trained on SUN RGB-D dataset using the AdamW [5] optimizer with the batch size of 16. The initial learning rate is 0.008 and weight decay is 0.01. The network is trained for 46 epochs. The decay steps are 24 and 32. imVoteNet [6] is a two-stage network, and we restore the parameters of stage 1 from the provided pre-trained model and train the network from stage 2. We implement our method on MMDetection3D [2] with one NVIDIA TITAN V GPU.

B. More Results

B.1. Quantitative Results

We show per-category results on ScanNet V2 and SUN RGB-D under different IoU thresholds. Table 1 shows the detailed results at mAP@0.5 for each object category in ScanNet V2 dataset. Note that mAP@0.5 is a fairly challenging metric as it basically requires more than 79% coverage in each dimension of a bounding box, we can get better performances on VoteNet [7] and its successors [1, 3, 9]. The results indicate that DisARM can help the backbones localize and recognize objects more accurately. As can be seen, we attain improvements on most categories against backbone methods, e.g. 12 of 18 categories on BRNet and 15 of 18 categories on GroupFree3D. For VoteNet [7], we obtain more than 10% improvements on most objects, including *cabinet*, *chair*, *table*, *door*, *window*, *bookshelf*, *counter*, *desk*, *curtain*, *refrigerator*, *shower curtain*, *sink*,

*Corresponding author: kevin.kai.xu@gmail.com

bathtub and *other furniture*. Our approach achieves 8.7 points improvement on a challenging category, i.e. *picture*, and achieves the best results on *window*, *picture*, *sink* and *bathtub* by leveraging the relation feature weighted by displacement. It is found that objects organized in common patterns, e.g. windows on the wall, bookshelves by the table, sinks placed in the washroom, usually get higher improvements, which demonstrates the effectiveness of displacement aware relation information.

Table 2 and Table 3 show the results of mAP@0.25 and mAP@0.5 on SUN RGB-D. In terms of mAP@0.25, we improve the performance on most categories, i.e. 8 of 10, and obtain the **state-of-the-art** performance by plugging DisARM into imVoteNet [6]. Compared to VoteNet [7], we improve the performance on 9 of 10 categories in terms of mAP@0.5, which demonstrates the superiority of our method. Furthermore, H3DNet [9] leverages 4 PointNet++ [8] in backbone to achieve the reported results on the SUN RGB-D dataset, while VoteNet [7] equipped with our lightweight DisARM outperforms the H3DNet [9] on mAP@0.5.

Notably, our method works well on both datasets, which indicates its outstanding generalization ability for different detection scenarios with different detectors.

In Figure 1, we show the PR curves of different methods on ScanNet V2 and SUN RGB-D respectively. Illustrated by the curves, our method performs better on IoU=0.5 which confirms that DisARM can improve detection accuracy of the backbone networks.

B.2. Qualitative Results

We also show more qualitative results of our method on ScanNet V2 and SUN RGB-D. The results of ScanNet V2 are shown in Figure 2, 3, 4. The figures demonstrate that our method achieves fairly better detection results against VoteNet [7].

The results of SUN RGB-D are shown in Figure 5, 6, 7. Compared to the BRNet [1], we also achieve comparable detection results.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8963–8972, 2021. 1, 2

[2] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020. 1

[3] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. *arXiv preprint arXiv:2104.00678*, 2021. 1

[4] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 1

[5] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[6] Charles R Qi, Xinlei Chen, Or Litany, and Leonidas J Guibas. Imvotenet: Boosting 3d object detection in point clouds with image votes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4404–4413, 2020. 1, 2, 3

[7] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 1, 2, 3, 4, 5, 6, 7, 8, 9

[8] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017. 2

[9] Zaiwei Zhang, Bo Sun, Haitao Yang, and Qixing Huang. H3dnet: 3d object detection using hybrid geometric primitives. In *European Conference on Computer Vision*, pages 311–329. Springer, 2020. 1, 2

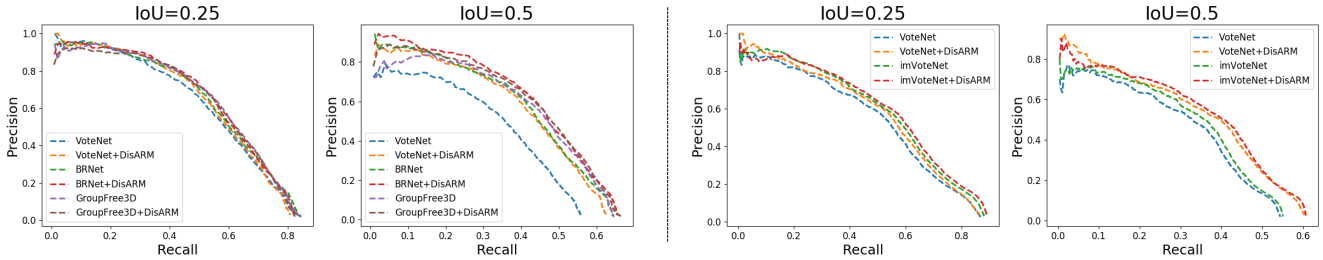

Figure 1. PR curves of backbone methods equipped with DisARM. The left two figures are the PR curves computed by IoU=0.25 and IoU=0.5 on ScanNet V2 dataset. The right two figures are the PR curves computed by IoU=0.25 and IoU=0.5 on SUN RGB-D dataset.

Method	cab	bed	chair	sofa	tabl	door	wind	bkshf	pic	cntr	desk	curt	fridg	showr	toil	sink	bath	ofurn	mAP
VoteNet	8.1	76.1	67.2	68.8	42.4	15.3	6.4	28.0	1.3	9.5	37.5	11.6	27.8	10.0	86.5	16.8	78.9	11.7	33.5
BRNet	**28.7**	80.6	81.9	80.6	60.8	35.5	22.2	48.0	7.5	**43.7**	54.8	39.1	**51.8**	35.9	88.9	38.7	84.4	33.0	50.9
H3DNet*	20.0	**85.5**	79.9	82.2	60.0	34.3	20.0	47.6	6.6	31.9	45.5	20.4	45.3	**52.9**	91.0	26.9	82.1	32.1	48.0
GroupFree3D^2	23.8	77.2	81.6	65.1	62.8	35.0	21.3	39.4	7.0	33.1	**66.3**	39.3	43.9	47.0	91.2	38.5	85.2	37.4	49.7
VoteNet*+DisARM	26.1	80.0	81.1	68.0	56.9	32.5	**27.5**	49.2	**10.0**	26.6	55.1	31.6	40.3	49.8	91.4	**44.5**	**89.5**	40.0	49.7
H3DNet*+DisARM	23.3	81.8	78.4	79.3	58.7	34.8	25.1	50.5	5.5	23.7	45.9	32.5	35.2	50.4	89.4	40.9	88.6	34.2	48.8
BRNet+DisARM	28.4	84.1	**85.4**	**83.9**	**62.9**	**38.9**	23.9	**55.8**	6.5	36.6	56.6	30.7	49.2	47.8	91.4	39.5	84.3	35.1	52.3
GroupFree3D^{2*}+DisARM	25.1	80.3	82.1	78.1	61.9	36.1	23.0	42.1	8.6	42.0	59.6	**48.4**	35.4	50.8	**94.5**	43.5	88.4	**44.4**	**52.5**

Table 1. 3D object detection results on ScanNet V2 dataset with mAP@0.5. * denotes the method is implemented with MMDetection3D. Note that VoteNet*+DisARM, H3DNet*+DisARM, BRNet+DisARM and GroupFree3D^{2*}+DisARM indicate applying our method to the 3D object detectors respectively. In addition, we mark GroupFree3D(L12, 0256) as GroupFree3D^2 and report the results in the official paper.

Method	bathtub	bed	bookshelf	chair	desk	dresser	nightstand	sofa	table	toilet	mAP
VoteNet	74.4	83.0	28.8	75.3	22.0	29.8	62.2	64.0	47.3	90.1	57.7
imVoteNet*	77.3	**88.2**	41.4	79.5	30.8	**39.9**	68.0	72.4	51.5	91.5	64.0
VoteNet*+DisARM	76.7	86.2	35.4	78.4	31.0	34.6	66.3	68.1	51.2	86.9	61.5
imVoteNet*+DisARM	**79.9**	87.5	**43.7**	**80.7**	**33.3**	39.8	**69.5**	**74.1**	**52.7**	**91.6**	**65.3**

Table 2. 3D object detection results on SUN RGD-D val dataset with mAP@0.25. * denotes the method is implemented with MMDetection3D. VoteNet*+DisARM and imVoteNet*+DisARM indicate applying our method to the VoteNet [7] and imVoteNet [6].

Method	bathtub	bed	bookshelf	chair	desk	dresser	nightstand	sofa	table	toilet	mAP
VoteNet*	45.4	53.4	6.8	56.5	5.9	12.0	38.6	49.1	21.3	68.5	35.8
H3DNet	47.6	52.9	8.6	60.1	8.4	20.6	45.6	50.4	**27.1**	**69.1**	39.0
VoteNet*+DisARM	**55.6**	**54.4**	**13.7**	**61.0**	**11.3**	**24.0**	**52.0**	**54.0**	26.9	60.0	**41.3**

Table 3. 3D object detection results on SUN RGD-D val dataset with mAP@0.5. * denotes the method is implemented with MMDetection3D and VoteNet*+DisARM indicates applying our method to the VoteNet [7]. Note that imVoteNet [6] does not provide the results in terms of mAP@0.5 for the SUN RGB-D dataset and we omit its comparison on this metric.

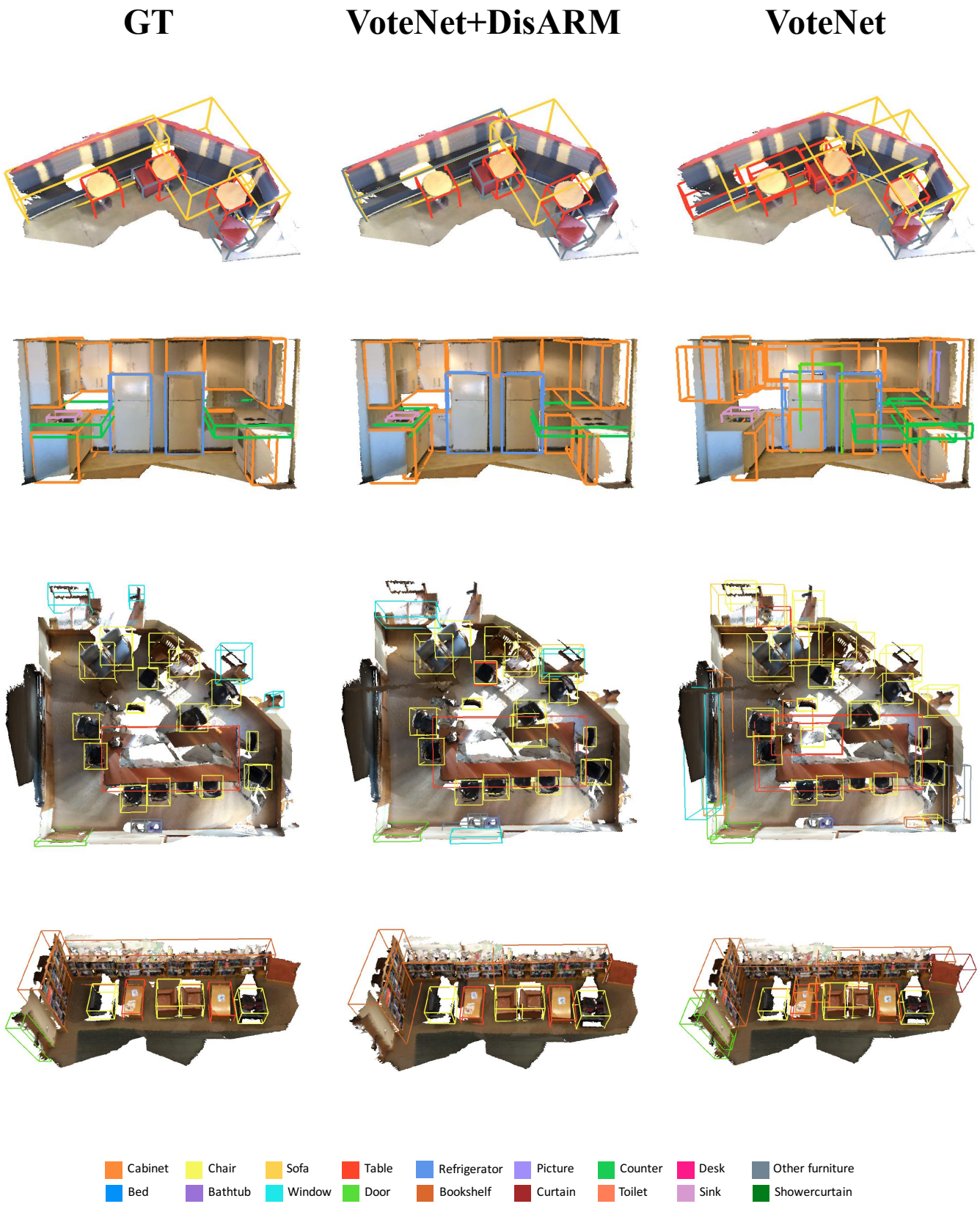

Figure 2. Qualitative results on ScanNet V2 dataset. We denote VoteNet+DisARM as applying our method to VoteNet [7]. The first column is ground truth. Best viewed on screen.

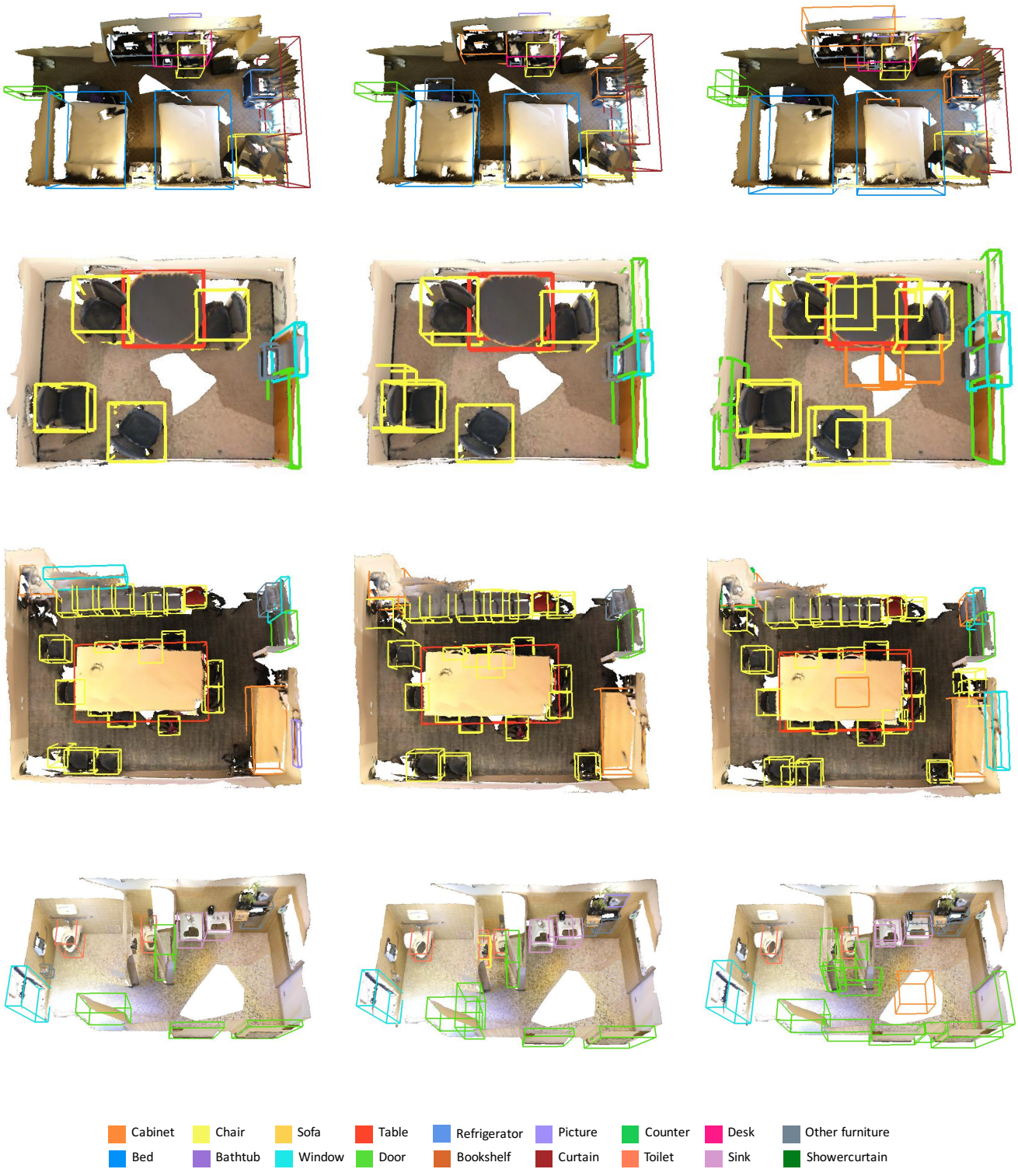

Figure 3. Qualitative results on ScanNet V2 dataset. We denote VoteNet+DisARM as applying our method to VoteNet [7]. The first column is ground truth. Best viewed on screen.

Figure 4. Qualitative results on ScanNet V2 dataset. We denote VoteNet+DisARM as applying our method to VoteNet [7]. The first column is ground truth. Best viewed on screen.

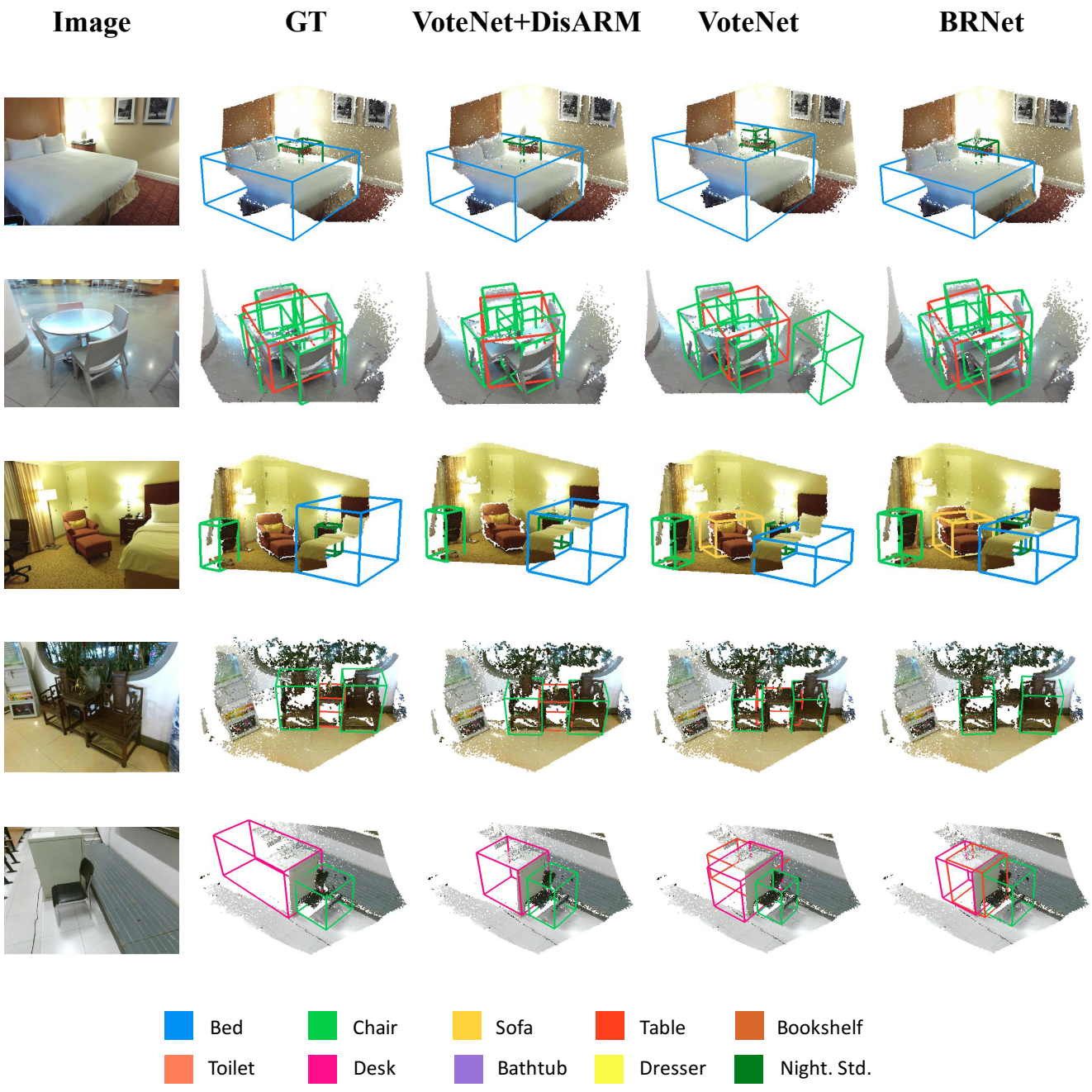

Figure 5. Qualitative results on SUN RGB-D dataset. We denote VoteNet+DisARM as applying our method to VoteNet [7]. The first column is ground truth and the rest columns are detections of different methods. Best viewed on screen.

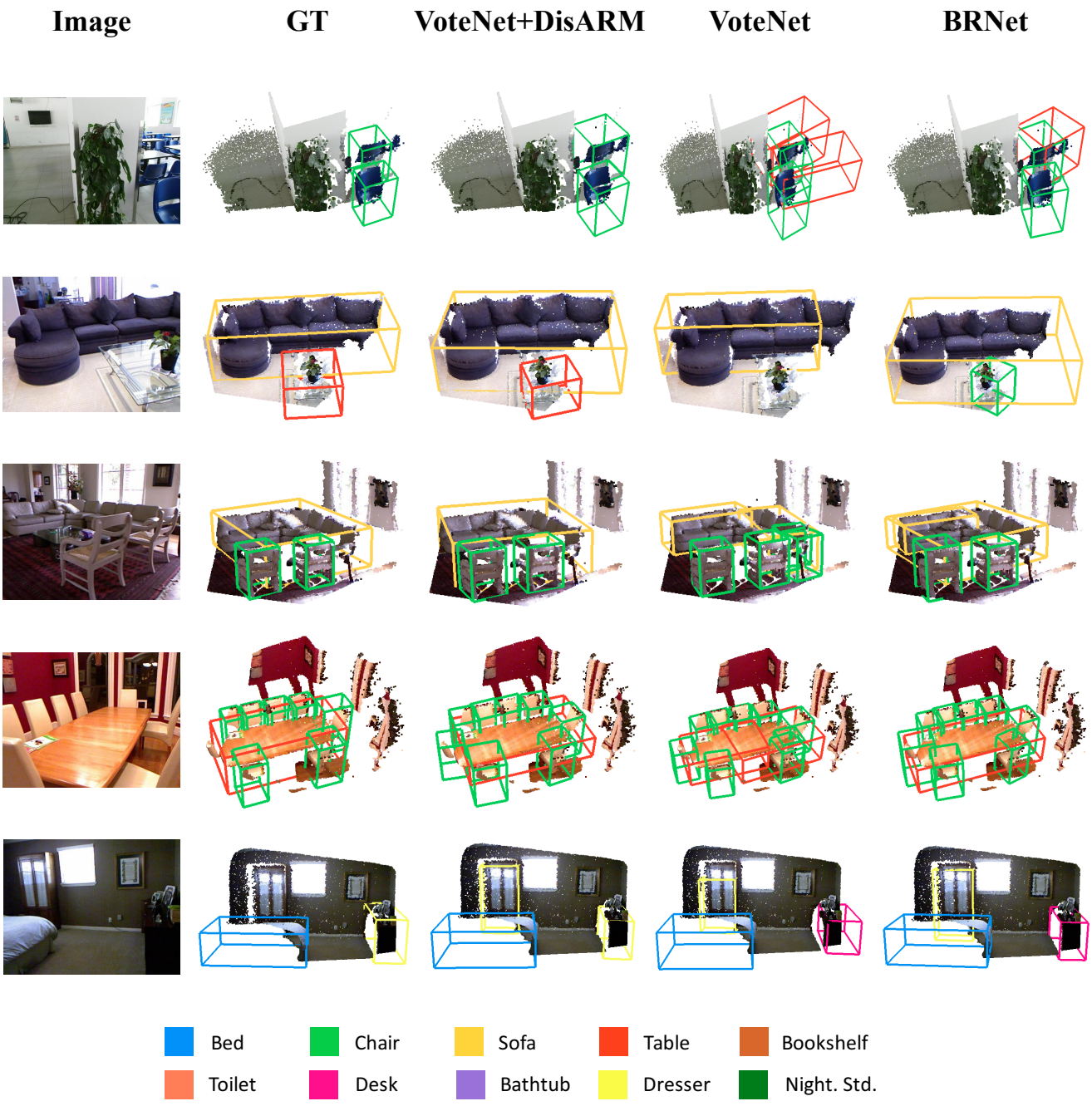

Figure 6. Qualitative results on SUN RGB-D dataset. We denote VoteNet+DisARM as applying our method to VoteNet [7]. The first column is ground truth and the rest columns are detections of different methods. Best viewed on screen.

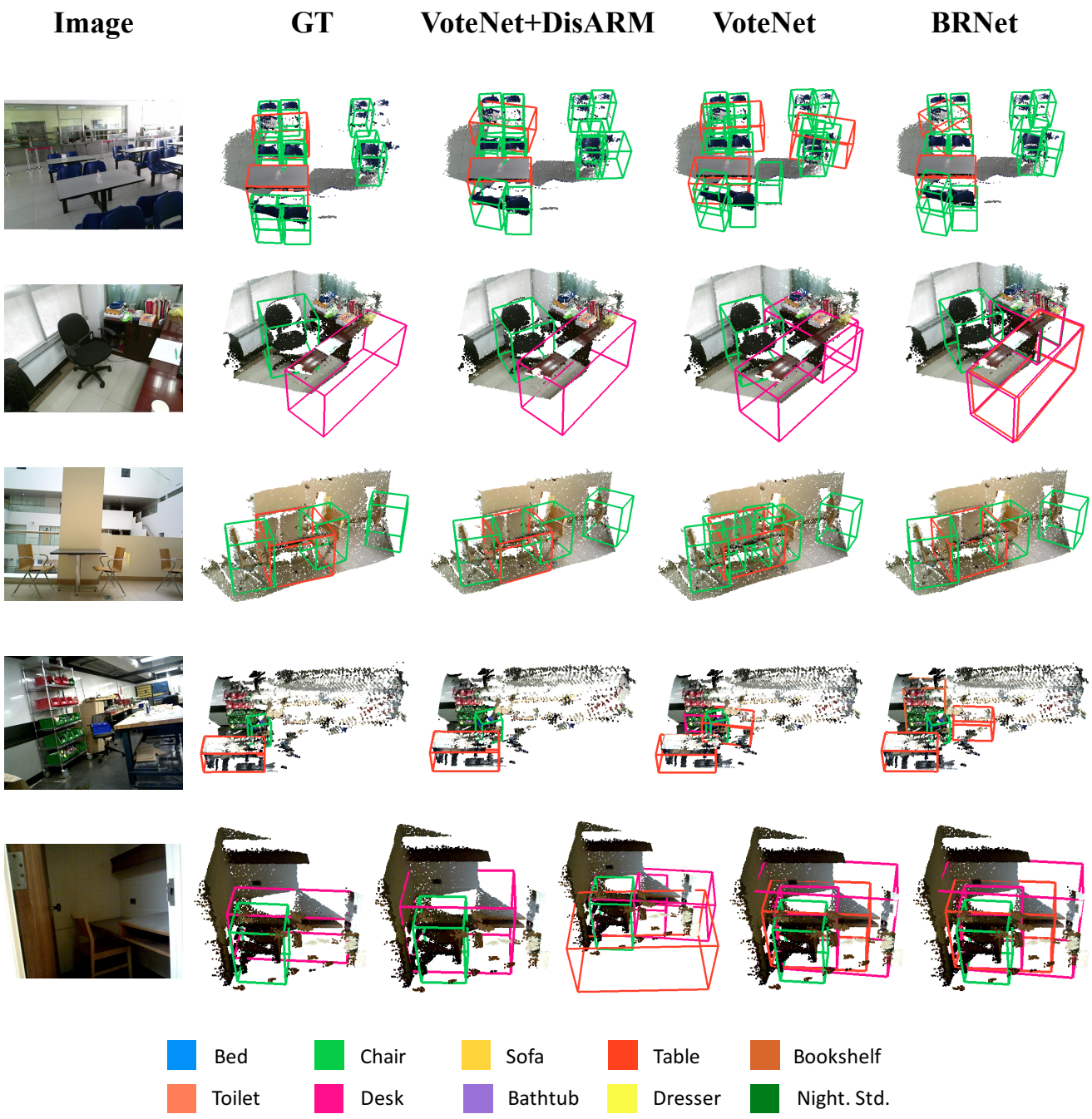

Figure 7. Qualitative results on SUN RGB-D dataset. We denote VoteNet+DisARM as applying our method to VoteNet [7]. The first column is ground truth and the rest columns are detections of different methods. Best viewed on screen.