ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

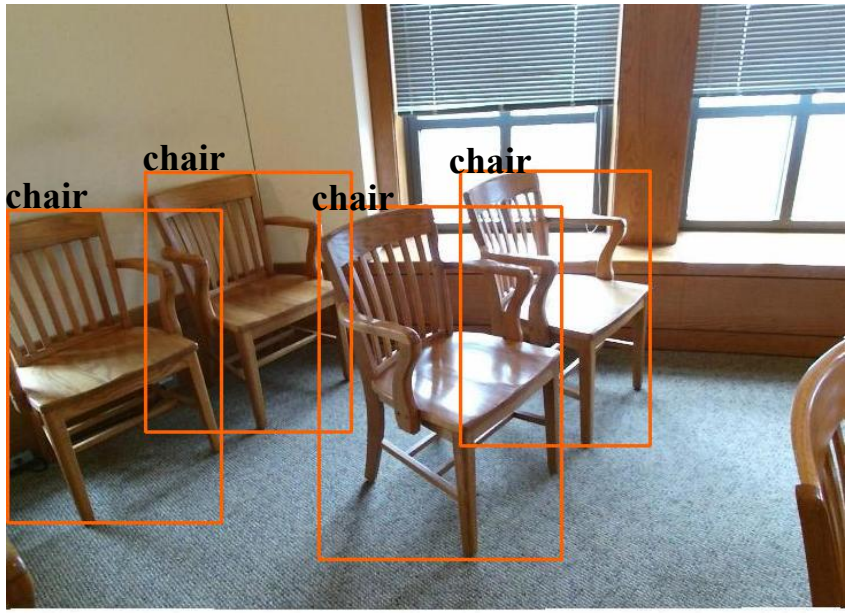
S-VoteNet: Deep Hough Voting with Spherical Proposal for 3D Object Detection

Yanxian Chen[1], Huimin Ma[1*], Xi Li[2], Xiong Luo[1]

[1]University of Science and Technology Beijing

[2]Tsinghua University

Introduction



2D Object Detection

Target: (x, y, w, h) + class

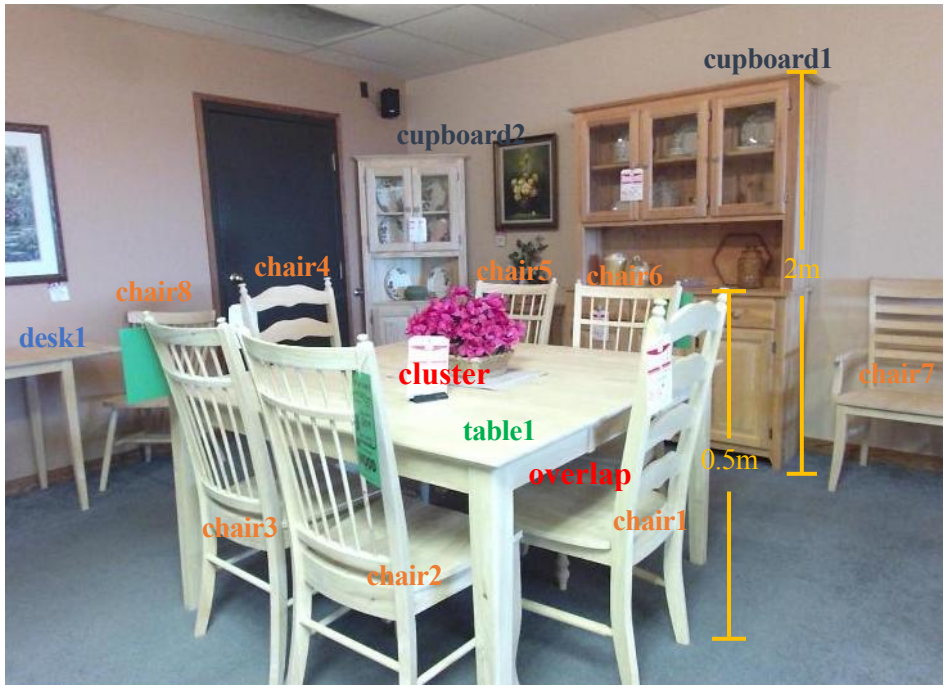
3D Object Detection

Target: (x, y, z, l, w, h, orientation) + class

- More parameters need to be predicted for 3D object detection.
- The orientation of objects has a great influence on the calculation of 3D IoU.

Introduction

Challenge of Indoor 3D Object Detection

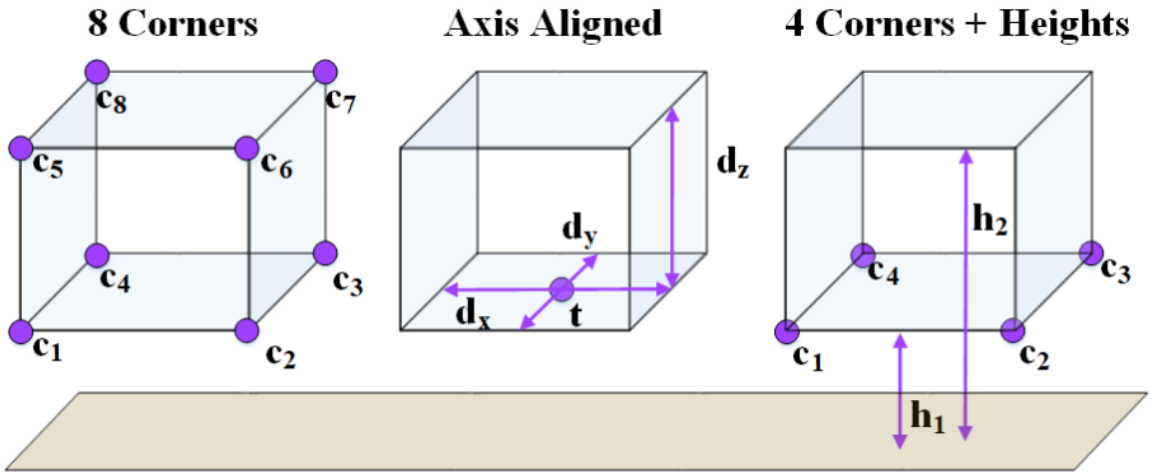


- Large variety and number of objects
- Large size differences between objects
- Complex spatial positions of objects

- In cluttered indoor scenes, it is difficult for 3d detectors to accurately predict the location and size of objects at the same time.

Related Work

3D Box Encoding



MV3D[1] AVOD[2]

- Axis Aligned 3D box:

  no orientation, (x, y, z, dx, dy, dz)

- Oriented 3D box:

  original, $(x, y, z, l, w, h, orientation)$

  8-corners, (x_i, y_i, z_i), $i \in [1, 8]$

  4-corners, h_{top}, h_{bottom}, (x_i, y_i), $i \in [1, 4]$

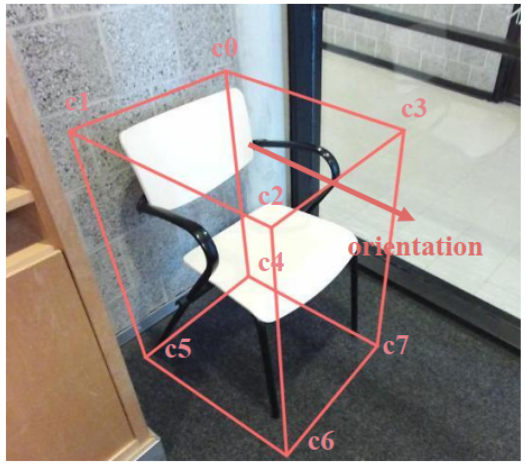
Object Location Loss

- $l2$ center loss: The **Euclidean distance** between proposal and ground truth is used as supervision.
- IoU loss: The **intersection over union**[3][4] between proposal and ground truth is used as supervision.

[1] Multi-view 3d object detection network for autonomous driving. Xiaozhi Chen, et al. CVPR 2017.
[2] Joint 3d proposal generation and object detection from view aggregation. Ku J, et al. IROS 2018.
[3] Generalized intersection over union: A metric and a loss for bounding box regression .Rezatofighi H, et al. CVPR 2019.
[4] Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. Zheng Z, Wang P, Liu W, et al. AAAI 2020.

Methodology

Spherical Encoding of 3D Box



8 corners encoding

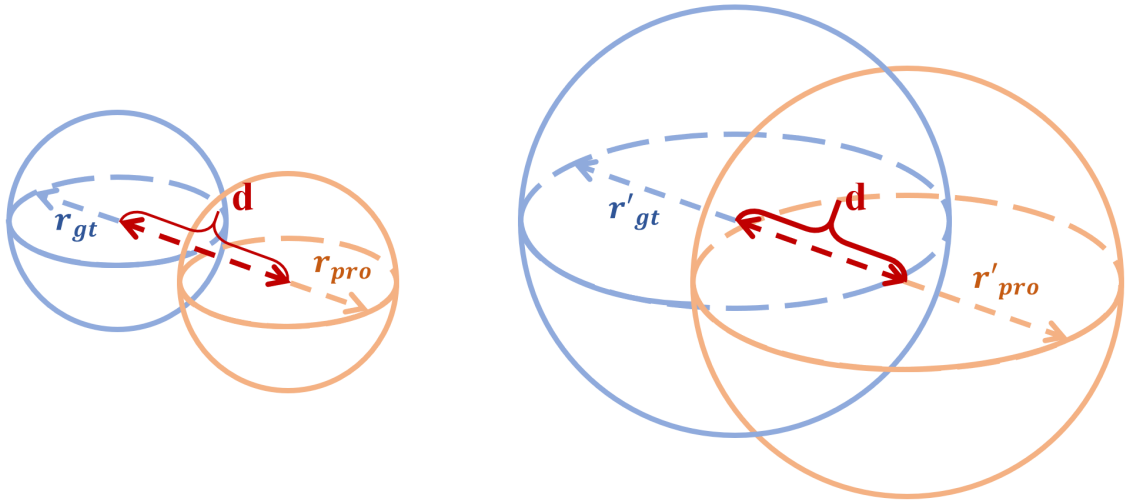
Spherical encoding

- Spherical encoding: (x, y, z, r)
- Based on spherical encoding, the 3D object detection task can be decoupled into the object location task and the size prediction task.
- For size and orientation prediction, We adopt method of F-PointNet[5].

[5] Frustum pointnets for 3d object detection from rgb-d data. Charles R Qi, et al. CVPR 2018

Methodology

Influence of Object Size on Object Location



(a) (b)

- ***l*2 center loss:**

$$L_{center} = d(c_{pro}, c_{gt})$$

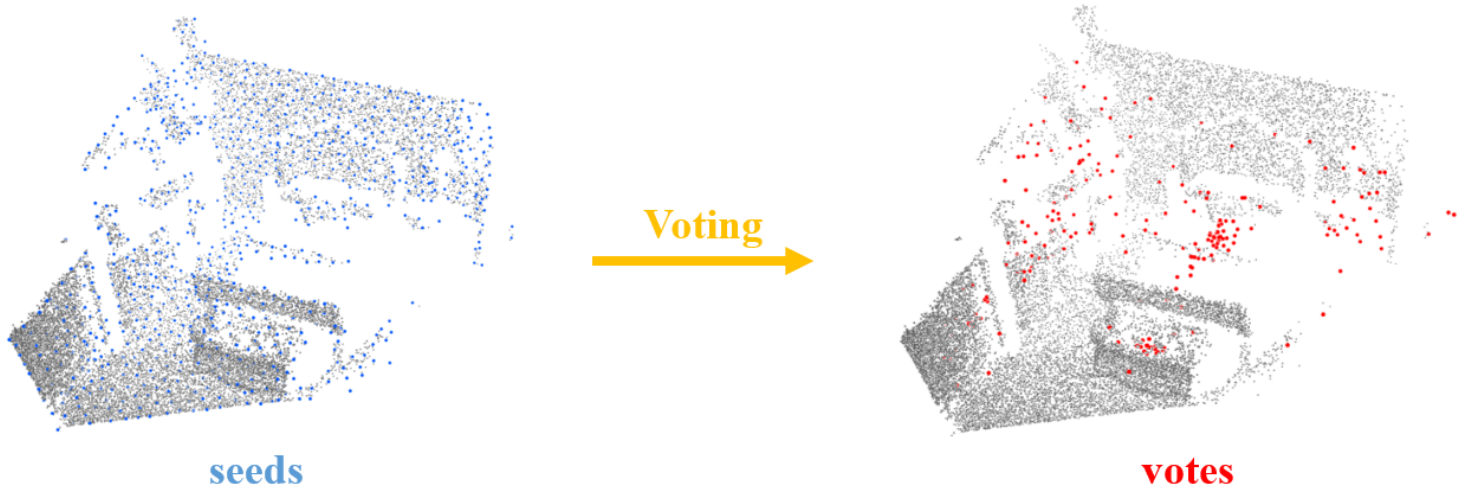
- **Spherical center loss:**

$$L_{spherical-center} = \frac{d(c_{pro}, c_{gt})}{d(c_{pro}, c_{gt}) + r_{pro} + r_{gt}}$$

- The distance between ground truth and proposal is the same for objects of different sizes, but the IoU is different.
- In this case, spherical center loss outputs adaptive localization loss based on object size, while *l*2 center loss does not.

Methodology

Geometric Information of Point Cloud:

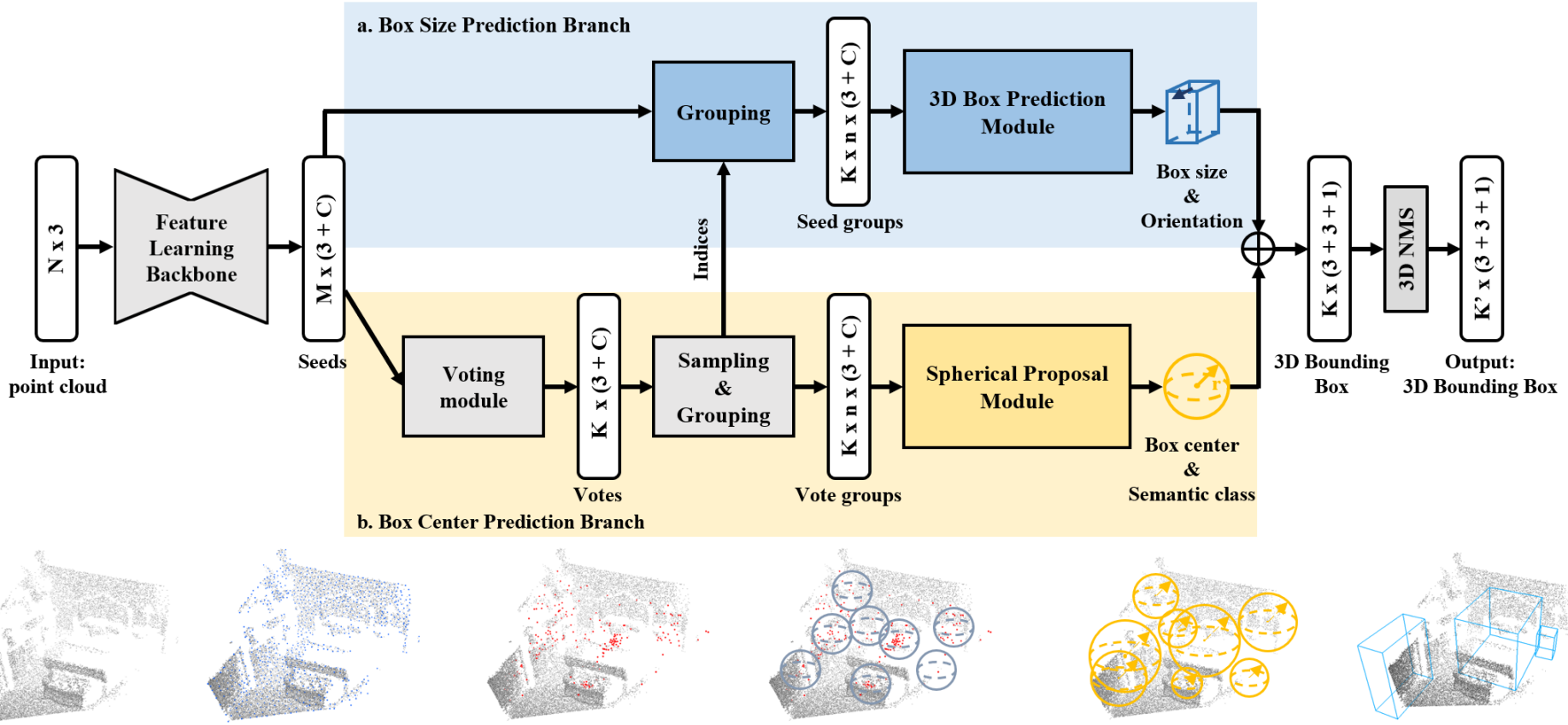


- Before voting, seeds preserve rich geometric information of the object.
- After voting[6], votes gather in the center of the object, which lose many of the geometric features.
- Seeds are suitable for object size and orientation prediction, while votes are fit for object location prediction.

[6] Deep hough voting for 3d object detection in point clouds. Charles R Qi, et al. ICCV 2019

Methodology

Overall Structure of S-VoteNet:



- S-VoteNet is built on the basis of VoteNet, which introduces spherical proposal to decouple the 3D object detection task.

[6] Deep hough voting for 3d object detection in point clouds. Charles R Qi, et al. ICCV 2019

Experiment

Performance on SUN RGB-D Val Set

TABLE I
PERFORMANCE COMPARISON OF 3D OBJECT DETECTION WITH PREVIOUS METHODS ON SUN RGB-D V1 VAL SET.

methods	input	bathtub	bed	bookshelf	chair	desk	dresser	nightstand	sofa	table	toilet	mAP
DSS [16]	Geo + RGB	44.2	78.8	11.9	61.2	20.5	6.4	15.4	53.5	50.3	78.9	42.1
COG [13]	Geo + RGB	58.3	63.7	31.8	62.2	**45.2**	15.5	27.4	51.0	**51.3**	70.1	47.6
2D-driven [4]	Geo + RGB	43.5	64.5	31.4	48.3	27.9	25.9	41.9	50.4	37.0	80.4	45.1
F-PointNet [8]	Geo + RGB	44.3	81.1	33.3	64.2	24.7	32.0	58.1	61.1	51.1	**90.9**	54.0
VoteNet + region feature [6]	Geo + RGB	71.7	86.1	34.0	74.7	26.0	34.2	64.3	66.5	49.7	88.4	59.6
ImVoteNet [6]	Geo + RGB	**75.9**	**87.6**	**41.3**	**76.7**	28.7	**41.4**	**69.9**	**70.7**	51.1	90.5	**63.4**
VoteNet [7]	Geo only	74.4	83.0	28.8	**75.3**	22.0	29.8	62.2	64.0	47.3	**90.1**	57.7
S-VoteNet (**ours**)	Geo only	**78.0**	**85.6**	**35.9**	74.6	**26.9**	**31.7**	**65.5**	**67.6**	**47.5**	89.3	**60.3**

- S-VoteNet advances the baseline by 2.6% mAP, which achieves performance second only to ImVoteNet without the use of RGB information.

Experiment

Analysis Study

TABLE II
EFFECTS OF VARIOUS COMPONENTS OF S-VOTENET

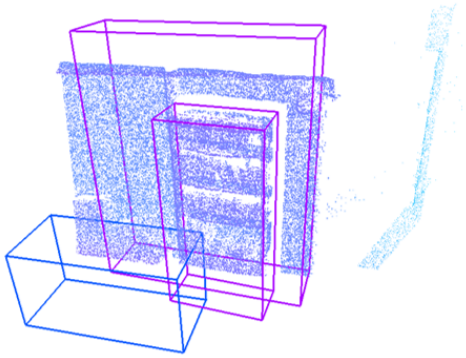
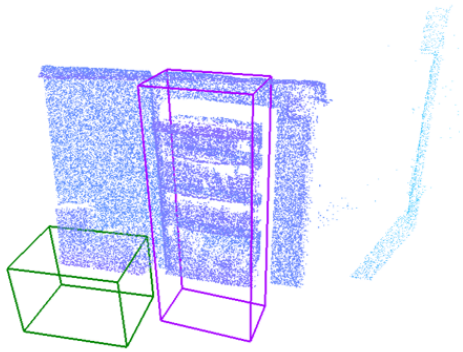
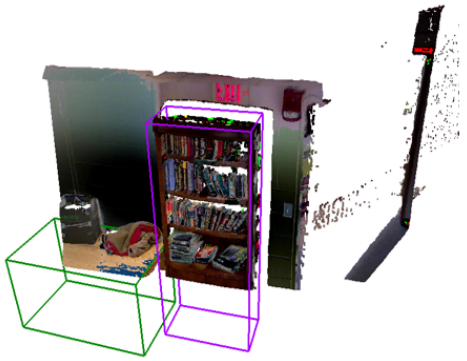
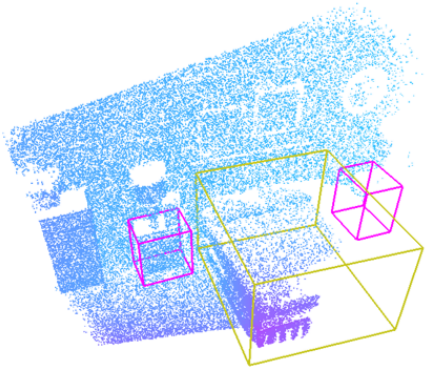
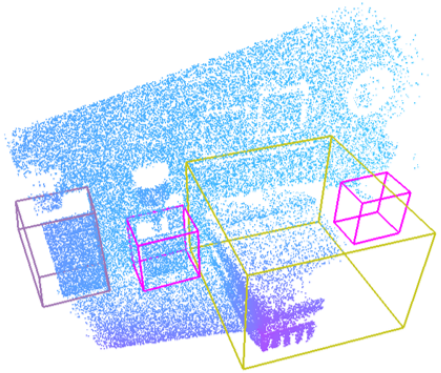
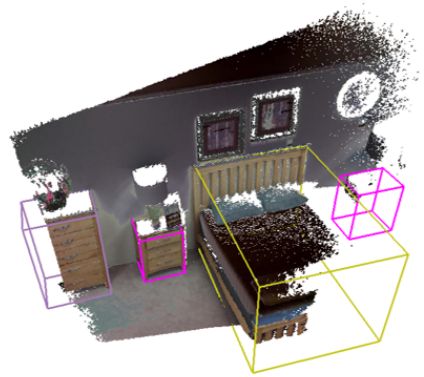
methods	use spherical center loss	use seed	mAP
BoxNet [7]	×	√	53.0
VoteNet [7]	×	×	57.7
VoteNet*	×	×	58.0
VoteNet**	√	×	59.5
S-VoteNet	√	√	60.3

- BoxNet is the baseline of VoteNet, which generates proposals without the voting module.
- VoteNet* is a variant of VoteNet, which decouples 3D object detection task without spherical encoding.
- VoteNet** is the improved version of VoteNet*, which introduces spherical center loss based on VoteNet*.
- S-VoteNet is the improved version of VoteNet**, which uses seeds to predict object size and orientation.

Experiment

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

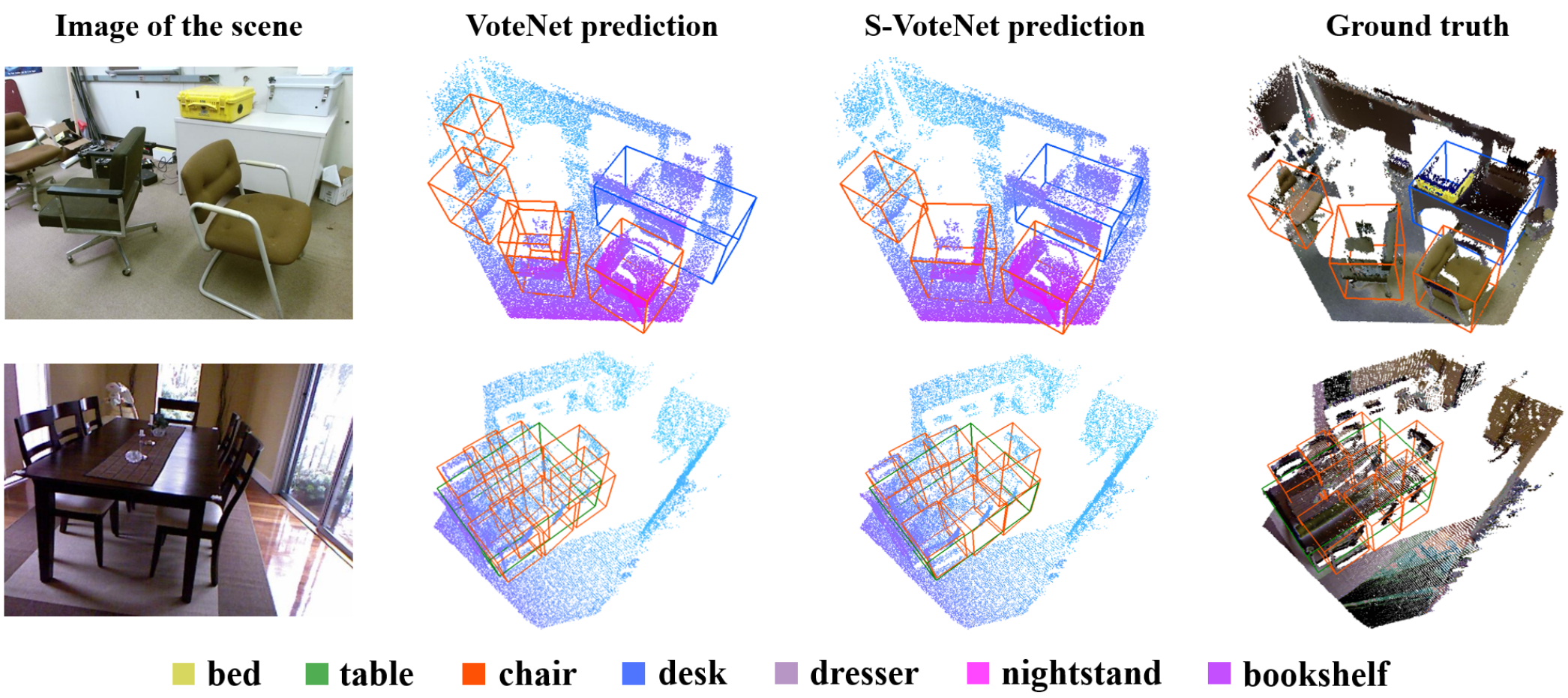
Qualitative results

Image of the scene	VoteNet prediction	S-VoteNet prediction	Ground truth
			
			

bed table chair desk dresser nightstand bookshelf

Experiment

Qualitative results



ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Thanks for Listening!