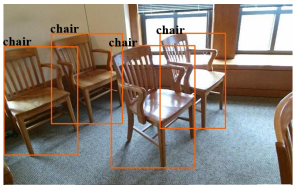
S-VoteNet: Deep Hough Voting with Spherical Proposal for 3D Object Detection

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

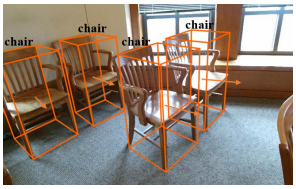
[1]University of Science and Technology Beijing [2]Tsinghua University

Motivation

Indoor Scene Object Detection:



2D Object Detection
$(x, y, w, h) + class$



3D Object Detection
$(x, y, z, l, w, h, orientation) + class$

Challenge of Indoor 3D Object Detection:

- Indoor scenes usually contain many objects, and the categories of objects are variety.
- There are large size differences between different categories of objects, which has a great impact on object localization.
- There are complex spatial relationships between objects, some clustered in groups, and some stacked on other objects.
- These factors make it difficult for 3D detectors to accurately predict the position and size of objects in a regression head.

Contribution:

- We show that with a proper structure, the decoupling of 3D bounding box regression can effectively improve the performance of the 3D detector.
- The proposed spherical center loss further considers the geometric distance between proposal and ground truth, which achieves higher 3D localization accuracy.
- Our S-VoteNet achieves state-of-the-art 3D object detection performance on SUN RGB-D dataset by only using point cloud as input.

Related Work

3D Box Encoding:

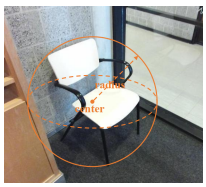
- Axis Aligned 3D box: no orientation, 6-dim vector, (x, y, z, dx, dy, dz)
- Oriented 3D box: original, 7-dim vector, $(x, y, z, l, w, h, orientation)$
  8-corners, 24-dim vector, $(x_i, y_i, z_i), i \in [1, 8]$
  4-corners, 10-dim vector, $h_{top}, h_{bottom}, (x_i, y_i), i \in [1, 4]$

Object Location Loss:

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

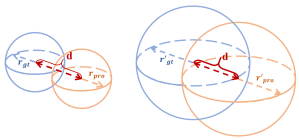
Methodology

Spherical Encoding:



- Spherical encoding: 4-dim vector, (x, y, z, r)
- Based on spherical encoding, 3D object detection task can be decoupled into object location task, size and orientation prediction task.
- For object location, we use spherical center loss to constrain the prediction result. For size and orientation, we adopt the method of F-PointNet.

Spherical Center Loss:

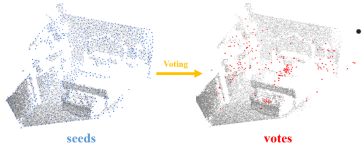


$$L_{center} = d$$

$$L_{spherical-center} = \frac{d}{d + r_{pro} + r_{gt}}$$

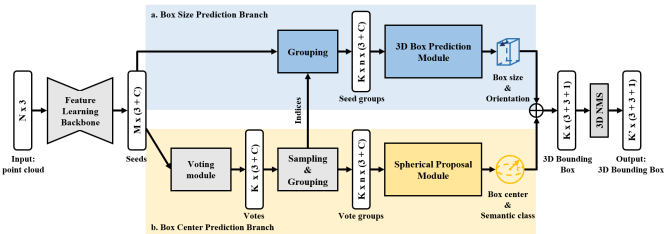
- The distance between ground truth and proposal is the same for objects of different sizes, but the IoU is different.
- Compared to $l2$ center loss, spherical center loss introduces object size into object center prediction, achieving higher 3D localization accuracy.

Geometric Information of Point Cloud:



- From the geometric information of point cloud on different stage, seeds are suitable for object size and orientation prediction, while votes are fit for object location prediction.

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.
- To align the object center predictions with the size and orientation predictions, we use votes indices to find the corresponding seeds.

Result

Performance on SUN RGB-D Val Set:

methods	input	bathtub	bed	bookshelf	chair	desk	dresser	nightstand	sofa	table	toilet	mAP
DSS	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	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	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	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	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	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	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.

Ablation Study:

methods	use spherical center loss	use seed	mAP
BoxNet	×	√	53.0
VoteNet	×	×	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.

Qualitative Results: