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

Yanxian Chen¹, Huimin Ma¹*, Xi Li², Xiong Luo¹
¹University of Science and Technology Beijing ²Tsinghua University

Motivation
Indoor Scene Object Detection:

2D Object Detection
\[(x, y, w, h) + \text{class}\]

3D Object Detection
\[(x, y, z, l, w, h, \text{orientation}) + \text{class}\]

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, \((x, y, z, l, w, h, \text{orientation})\)

8-corners, 24-dim vector, \((x_i, y_i, z_i) i \in \{1, 8\}\)
4-corners, 10-dim vector, \(h_{\text{top}}, h_{\text{bottom}}, (x_i, y_i) i \in \{1, 4\}\)

Object Location Loss:

- I2 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:

\[
L_{\text{center}} = d - r_{\text{pro}} - r_{\text{gt}}
\]

- The distance between ground truth and proposal is the same for objects of different sizes, but the IoU is different.
- Compared to I2 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:

<table>
<thead>
<tr>
<th>methods</th>
<th>input</th>
<th>bathtub</th>
<th>bed</th>
<th>bookshelf</th>
<th>chair</th>
<th>desk</th>
<th>dresser</th>
<th>nightstand</th>
<th>sofa</th>
<th>table</th>
<th>toilet</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSS</td>
<td>Geo + RGB</td>
<td>44.2</td>
<td>78.8</td>
<td>31.9</td>
<td>61.2</td>
<td>20.5</td>
<td>6.4</td>
<td>15.4</td>
<td>53.5</td>
<td>50.3</td>
<td>78.9</td>
<td>42.1</td>
</tr>
<tr>
<td>COG</td>
<td>Geo + RGB</td>
<td>58.3</td>
<td>63.7</td>
<td>31.8</td>
<td>62.2</td>
<td>45.2</td>
<td>15.5</td>
<td>27.4</td>
<td>51.0</td>
<td>53.3</td>
<td>70.1</td>
<td>47.6</td>
</tr>
<tr>
<td>2D-driven</td>
<td>Geo + RGB</td>
<td>43.5</td>
<td>64.5</td>
<td>31.4</td>
<td>48.3</td>
<td>27.9</td>
<td>25.9</td>
<td>41.9</td>
<td>50.4</td>
<td>37.0</td>
<td>80.4</td>
<td>45.1</td>
</tr>
<tr>
<td>F-PointNet</td>
<td>Geo + RGB</td>
<td>44.3</td>
<td>81.1</td>
<td>33.5</td>
<td>64.2</td>
<td>42.7</td>
<td>32.0</td>
<td>58.1</td>
<td>61.1</td>
<td>51.1</td>
<td>98.9</td>
<td>54.0</td>
</tr>
<tr>
<td>VoteNet</td>
<td>Geo + RGB</td>
<td>71.7</td>
<td>83.1</td>
<td>34.0</td>
<td>74.3</td>
<td>26.0</td>
<td>34.2</td>
<td>64.3</td>
<td>66.5</td>
<td>49.7</td>
<td>88.4</td>
<td>59.6</td>
</tr>
<tr>
<td>ImVoteNet</td>
<td>Geo + RGB</td>
<td>75.9</td>
<td>87.6</td>
<td>41.3</td>
<td>76.7</td>
<td>28.7</td>
<td>41.4</td>
<td>69.9</td>
<td>70.7</td>
<td>51.1</td>
<td>99.5</td>
<td>63.4</td>
</tr>
</tbody>
</table>

- VoteNet only
- S-VoteNet (ours) Geo only

<table>
<thead>
<tr>
<th>methods</th>
<th>use spherical center loss</th>
<th>use seed</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoxNet</td>
<td>✓</td>
<td></td>
<td>53.0</td>
</tr>
<tr>
<td>VoteNet</td>
<td>✓</td>
<td></td>
<td>57.7</td>
</tr>
<tr>
<td>VoteNet*</td>
<td>✓</td>
<td></td>
<td>58.0</td>
</tr>
<tr>
<td>VoteNet**</td>
<td>✓</td>
<td></td>
<td>59.5</td>
</tr>
<tr>
<td>S-VoteNet</td>
<td>✓</td>
<td></td>
<td>60.3</td>
</tr>
</tbody>
</table>

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

Ablation Study:

- 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 a variant of VoteNet, which decouples 3D object detection task without spherical encoding.