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
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

- Axis Aligned 3D box: no orientation, \((x, y, z, dx, dy, dz)\)
- Oriented 3D box: original, \((x, y, z, l, w, h, \text{orientation})\)
  - 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

- \(l_2\) 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.

---

Methodology

Spherical Encoding of 3D Box

- 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].

Methodology

Influence of Object Size on Object Location

- **$l_2$ center loss:**
  \[
  L_{\text{center}} = d(c_{\text{pro}}, c_{\text{gt}})
  \]

- **Spherical center loss:**
  \[
  L_{\text{spherical-center}} = \frac{d(c_{\text{pro}}, c_{\text{gt}})}{d(c_{\text{pro}}, c_{\text{gt}}) + r_{\text{pro}} + r_{\text{gt}}}
  \]

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

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.

**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.

<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 [16]</td>
<td>Geo + RGB</td>
<td>44.2</td>
<td>78.8</td>
<td>11.9</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 [13]</td>
<td>Geo + RGB</td>
<td>58.3</td>
<td>63.7</td>
<td>31.8</td>
<td>62.2</td>
<td>15.5</td>
<td>27.4</td>
<td>51.0</td>
<td>51.3</td>
<td>70.1</td>
<td>47.6</td>
<td>45.1</td>
</tr>
<tr>
<td>2D-driven [4]</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 [8]</td>
<td>Geo + RGB</td>
<td>44.3</td>
<td>81.1</td>
<td>33.3</td>
<td>64.2</td>
<td>24.7</td>
<td>32.0</td>
<td>58.1</td>
<td>61.1</td>
<td>51.1</td>
<td>90.9</td>
<td>54.0</td>
</tr>
<tr>
<td>VoteNet + region feature [6]</td>
<td>Geo + RGB</td>
<td>71.7</td>
<td>86.1</td>
<td>34.0</td>
<td>74.7</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 [6]</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>90.5</td>
<td>63.4</td>
</tr>
<tr>
<td>VoteNet [7]</td>
<td>Geo only</td>
<td>74.4</td>
<td>83.0</td>
<td>28.8</td>
<td>75.3</td>
<td>22.0</td>
<td>29.8</td>
<td>62.2</td>
<td>64.0</td>
<td>47.3</td>
<td>90.1</td>
<td>57.7</td>
</tr>
<tr>
<td>S-VoteNet (ours)</td>
<td>Geo only</td>
<td>78.0</td>
<td>85.6</td>
<td>35.9</td>
<td>74.6</td>
<td>26.9</td>
<td>31.7</td>
<td>65.5</td>
<td>67.6</td>
<td>47.5</td>
<td>89.3</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.
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

Qualitative results

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

- bed
- table
- chair
- desk
- dresser
- nightstand
- bookshelf
<table>
<thead>
<tr>
<th>Experiment</th>
</tr>
</thead>
</table>

Qualitative results

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

Colors:  
- yellow = bed  
- green = table  
- orange = chair  
- blue = desk  
- purple = dresser  
- pink = nightstand  
- brown = bookshelf
Thanks for Listening!