Supplementary Materials for
HyperDet3D: Learning a Scene-conditioned 3D Object Detector

A. Technical Details

Here we introduce some technical details on the proposed HyperDet3D, including the processing of obtaining \( b \), and some hyper-parameters for the experiments on 2 datasets.

A.1. Processing of obtaining biases

For bias \( b \), we maintain \( Z^* \) and \( Z^s \) which are the same as in obtaining \( W \). The scene-specific and scene-agnostic hypernetworks both contain a single linear layer. Before the second attention, we take the average results of scene-specific and scene-agnostic vectors along the feature dimension.

A.2. Hyper-parameters

Following [3], the number of decoder layers for ScanNet [2] and SUN RGB-D [5] are 12 and 6 respectively. We set \( C_a=256 \) and \( n=256 \) for both datasets. We set \( C_s=253 \) and 285 for ScanNet and SUN RGB-D respectively.

B. Other Experimental Results

B.1. Detailed Results on SUN RGB-D

As shown in Table 1, we detail the per-category results on the SUN RGB-D [5] dataset. We observe the obvious improvements on category bed, bookshelf, nightstand and toilet which are more conditioned on its corresponding scenes (bedroom, washroom, etc.). The exception is the bathtub category with performance drop of -4.9%, which might be attributed to the extremely scarcity of its annotations in the dataset (the least one in the statistical distribution shown in the paper [5]).

B.2. Detailed Results on Cross-dataset Evaluation

For the cross-dataset experiments where we pretrained HyperDet3D on SUN RGB-D and finetuned on ScanNet v2, we display the detailed per-category results in Table 2, including the extra 10 novel categories.

B.3. Ablation Study on Cross-dataset Evaluation

The removal of scene-specific knowledge and MSA→SSA brings -3.0% and -3.4% on mAP\(_s\) respectively. We infer the expressiveness of MSA contributes more than scene-specific knowledge when tackling domain gap.

C. More Visualization Results

We have additionally illustrated qualitative results on ScanNet v2 and SUN RGB-D. The results on ScanNet v2 are shown in Figure 1. The results on SUN RGB-D are shown in Figure 2.

References

Table 1. 3D object detection results on SUN RGB-D V1 validation dataset. We show per-category results of mean average precision (mAP) with 3D IoU threshold 0.5 as proposed in [5], and mean of AP across all semantic classes with 3D IoU threshold 0.5.

<table>
<thead>
<tr>
<th></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>Votenet [4]</td>
<td>45.4</td>
<td>53.4</td>
<td>6.8</td>
<td>56.5</td>
<td>5.9</td>
<td>12.0</td>
<td>38.6</td>
<td>49.1</td>
<td>21.3</td>
<td>68.5</td>
<td>35.8</td>
</tr>
<tr>
<td>H3DNet [6]</td>
<td>47.6</td>
<td>52.9</td>
<td>8.6</td>
<td>60.1</td>
<td>8.4</td>
<td>20.6</td>
<td>45.6</td>
<td>50.4</td>
<td>27.1</td>
<td>69.1</td>
<td>39.0</td>
</tr>
<tr>
<td>BRNet [1]</td>
<td>55.5</td>
<td>63.8</td>
<td>9.3</td>
<td>61.6</td>
<td>10.0</td>
<td>27.3</td>
<td>53.2</td>
<td>56.7</td>
<td>28.6</td>
<td>70.9</td>
<td>43.7</td>
</tr>
<tr>
<td>GF3D [3]</td>
<td>64.0</td>
<td>67.1</td>
<td>12.4</td>
<td>62.6</td>
<td>14.5</td>
<td>21.9</td>
<td>49.8</td>
<td>58.2</td>
<td>29.2</td>
<td>72.2</td>
<td>45.2</td>
</tr>
<tr>
<td>Ours</td>
<td>59.1</td>
<td>69.4</td>
<td>21.1</td>
<td>62.1</td>
<td>11.3</td>
<td>14.0</td>
<td>57.4</td>
<td>61.3</td>
<td>27.6</td>
<td>89.8</td>
<td>47.3</td>
</tr>
</tbody>
</table>

Table 2. 3D object detection results on the ScanNet V2 validation dataset. We show per-category results of mean average precision (mAP) with 3D IoU threshold 0.5 as proposed in [5], and mean of AP across all semantic classes with 3D IoU threshold 0.5.

|            | cab     | bed    | chair   | sofa   | tabl   | door   | wind    | bkshf  | pic    | ctttr  | desk   | ctt   | curt   | ftdg   | showe  | toil   | sink   | bath   | ofurn  | mAP   |
|------------|---------|--------|---------|--------|--------|--------|---------|--------|--------|--------|--------|-------|-------|--------|--------|--------|--------|--------|--------|-------|-------|
| GF3D [3]   | 5.3     | 66.6   | 21.3    | 46.9   | 17.8   | 3.3    | 4.5     | 0.4    | 0      | 8.0    | 25.6   | 6.7   | 22.3   | 0      | 54.6   | 11.7   | 48.6   | 0.4    | 19.1  |
| Ours       | 10.8    | 78.9   | 22.7    | 58.0   | 16.0   | 2.9    | 2.2     | 2.4    | 0.5    | 13.3   | 40.1   | 11.6  | 7.4    | 0      | 58.9   | 0.5    | 71.4   | 2.3    | 22.2  |

Figure 1. Qualitative detection results on the ScanNet V2 validation set. (Best viewed in color.)
Figure 2. Qualitative detection results on the SUN RGB-D V1 validation set. (*Best viewed in color.*)