Supplementary Material:
HybridCR: Weakly-Supervised 3D Point Cloud Semantic Segmentation via Hybrid Contrastive Regularization

Mengtian Li\textsuperscript{1}, Yuan Xie\textsuperscript{1}, Yunhang Shen\textsuperscript{2}, Bo Ke\textsuperscript{2}, Ruizhi Qiao\textsuperscript{2}, Bo Ren\textsuperscript{2}, Shaohui Lin\textsuperscript{1,†}, Lizhuang Ma\textsuperscript{1,†}
\textsuperscript{1}School of Computer Science and Technology
East China Normal University, Shanghai, China
\textsuperscript{2}Tencent Youtu Lab
mtli@stu.ecnu.edu.cn, \{yxie, shlin, lzma\}@cs.ecnu.edu.cn
\{odysseyshen, boke, ruizhiqiao, timren\}@tencent.com

1. Overview

In the supplementary material, we start with more details of the training setup in Sec. 1.1 and model complexity in Sec. 1.2. Further, we give the per-class scores of ScanNet-V2 \cite{4}, Semantic3D \cite{6} and SemanticKITTI \cite{2} in Sec. 1.3, and provide more visual results in Sec. 1.4. Finally, we compare the dynamic augmentor and the fixed one in Sec. 1.5.

1.1. Training Setup

**Weakly Setting.** We create the weakly-supervised dataset by randomly annotating a tiny fraction of points in a class for each point cloud sample. Specifically, we set up two weakly-supervised training methods: 1pt and 1%. At 1pt setting, we annotate one point for each class for each point cloud sample. At 1% setting, we select 1% of the points that are labeled for each class randomly, and these labeled points will not change during the training. Thus, at the semantic level, we only annotate some points for each semantic class as this is a form of weak-supervision (incomplete supervision) defined by Zhou \textit{et al}. \cite{20}. In addition, Xu \textit{et al}. \cite{17} and Zhang \textit{et al}. \cite{18} also define incomplete supervision as a weakly-supervised task. Therefore, we follow the definition in this paper.

**Training configuration.** Here we have supplemented the experimental details of the main paper. Our network training is conducted on the RTX Titan GPU with 24 GB memory. We use a grid size of 4cm for indoor dataset and 6cm for outdoor dataset to down-sample the raw point clouds, let the barycenter of each small grid be the selected point. Then the network takes input point clouds of size 40960 points for all datasets during training.

<table>
<thead>
<tr>
<th>Method</th>
<th>Training time</th>
<th>Network parameters</th>
<th>Total reference time</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSD(1%) \cite{18}</td>
<td>302</td>
<td>1.10</td>
<td>263</td>
</tr>
<tr>
<td>HybridCR(1%)</td>
<td>387</td>
<td>1.51</td>
<td>279</td>
</tr>
</tbody>
</table>

Table 1. The training time of per-epoch (in seconds), the network parameters (in millions) and total test time (in seconds) on S3DIS.

1.2. Model Complexity

We list the training time of per-epoch, the network parameters, and the total test time in Tab. 1 compared with PSD \cite{18}. Since the parameters of the Siamese network are shared, only the parameters of dynamic point cloud augmentor are added compared to the PSD, so that the parameters of HybridCR are more by 0.41M than PSD. Since the augmentation operation and pseudo label selection are only introduced in the training phase, the training time of HybridCR is 85s per epoch longer than PSD. In comparison, the total reference time of HybridCR is relatively similar with PSD. Considering the significant improvements on quantitative results provided by HybridCR, it is still an efficient method.

1.3. Detailed Quantitative Results

**Evaluation on ScanNet-V2.** We present the segmentation performance of per-class on the ScanNet-V2 and choose the weakly-supervised setting of 1% for comparison. From Tab. 2. It can be observed that our HybridCR achieves 56.8% mIoU and 2.1% improvements against PSD. Moreover, in the aspect of specific classes, our method gains 11.1%, 8.8%, 8.4%, improvements in “door”, “other-furniture”, “curtain” against PSD, respectively, and achieve the best performance on “picture”. While HybridCR can significantly improve the performance of these classes and demonstrate that our method can learn more discrimina-
Table 2. Quantitative results of per class on ScanNet-V2 [4]. (mIoU %)

<table>
<thead>
<tr>
<th>Set</th>
<th>Methods</th>
<th>bath-tub</th>
<th>bed</th>
<th>bookshelf</th>
<th>cabinet</th>
<th>chair</th>
<th>curtain</th>
<th>desk</th>
<th>door</th>
<th>floor</th>
<th>other-furniture</th>
<th>picture</th>
<th>refrigerator</th>
<th>shower-curtain</th>
<th>sink</th>
<th>sofa</th>
<th>table</th>
<th>toilet</th>
<th>wall</th>
<th>window</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully</td>
<td>PointNet [12]</td>
<td>33.9</td>
<td>58.4</td>
<td>47.8</td>
<td>45.8</td>
<td>25.6</td>
<td>36.0</td>
<td>25.0</td>
<td>24.7</td>
<td>27.8</td>
<td>26.1</td>
<td>67.7</td>
<td>18.3</td>
<td>11.7</td>
<td>21.2</td>
<td>14.5</td>
<td>36.4</td>
<td>34.6</td>
<td>23.2</td>
<td>54.8</td>
</tr>
<tr>
<td></td>
<td>PCNN [1]</td>
<td>59.8</td>
<td>55.9</td>
<td>64.4</td>
<td>56.0</td>
<td>42.0</td>
<td>71.1</td>
<td>22.9</td>
<td>41.4</td>
<td>43.6</td>
<td>35.2</td>
<td>94.1</td>
<td>32.4</td>
<td>15.5</td>
<td>23.8</td>
<td>38.7</td>
<td>49.3</td>
<td>52.9</td>
<td>50.9</td>
<td>81.3</td>
</tr>
<tr>
<td></td>
<td>SegGCN [8]</td>
<td>58.9</td>
<td>83.3</td>
<td>73.1</td>
<td>53.9</td>
<td>51.4</td>
<td>78.9</td>
<td>44.8</td>
<td>46.7</td>
<td>57.3</td>
<td>48.4</td>
<td>93.6</td>
<td>39.6</td>
<td>6.1</td>
<td>50.1</td>
<td>50.7</td>
<td>59.4</td>
<td>70.0</td>
<td>56.3</td>
<td>87.4</td>
</tr>
<tr>
<td></td>
<td>PointConv [16]</td>
<td>66.6</td>
<td>78.1</td>
<td>75.9</td>
<td>69.9</td>
<td>64.4</td>
<td>82.2</td>
<td>47.5</td>
<td>77.9</td>
<td>56.4</td>
<td>50.4</td>
<td>95.3</td>
<td>42.8</td>
<td>20.3</td>
<td>58.6</td>
<td>75.4</td>
<td>66.1</td>
<td>75.3</td>
<td>58.8</td>
<td>90.2</td>
</tr>
<tr>
<td></td>
<td>KPConv [14]</td>
<td>68.4</td>
<td>84.7</td>
<td>75.8</td>
<td>78.4</td>
<td>64.7</td>
<td>81.4</td>
<td>47.3</td>
<td>77.2</td>
<td>60.5</td>
<td>59.4</td>
<td>93.5</td>
<td>45.0</td>
<td>18.1</td>
<td>58.7</td>
<td>80.5</td>
<td>69.0</td>
<td>78.5</td>
<td>61.4</td>
<td>88.2</td>
</tr>
<tr>
<td></td>
<td>RFCR [5]</td>
<td>70.2</td>
<td>88.9</td>
<td>74.5</td>
<td>81.3</td>
<td>67.2</td>
<td>81.8</td>
<td>49.3</td>
<td>81.5</td>
<td>62.3</td>
<td>61.0</td>
<td>94.7</td>
<td>47.0</td>
<td>24.9</td>
<td>59.4</td>
<td>84.8</td>
<td>70.5</td>
<td>77.9</td>
<td>64.6</td>
<td>89.2</td>
</tr>
<tr>
<td></td>
<td>HybridCR</td>
<td>59.9</td>
<td>87.2</td>
<td>70.7</td>
<td>68.3</td>
<td>56.1</td>
<td>78.4</td>
<td>46.3</td>
<td>61.6</td>
<td>46.5</td>
<td>45.6</td>
<td>93.6</td>
<td>42.7</td>
<td>20.7</td>
<td>46.4</td>
<td>56.7</td>
<td>53.1</td>
<td>69.5</td>
<td>48.0</td>
<td>71.3</td>
</tr>
<tr>
<td>weakly</td>
<td>PSD(1%) [18]</td>
<td>54.7</td>
<td>57.1</td>
<td>67.8</td>
<td>65.9</td>
<td>46.5</td>
<td>77.8</td>
<td>38.8</td>
<td>52.8</td>
<td>49.2</td>
<td>30.4</td>
<td>93.3</td>
<td>38.7</td>
<td>30.7</td>
<td>43.1</td>
<td>38.2</td>
<td>52.6</td>
<td>66.9</td>
<td>57.2</td>
<td>71.6</td>
</tr>
<tr>
<td></td>
<td>HybridCR(1%)</td>
<td>56.8</td>
<td>58.9</td>
<td>65.8</td>
<td>66.8</td>
<td>42.3</td>
<td>80.2</td>
<td>36.7</td>
<td>61.2</td>
<td>58.1</td>
<td>45.5</td>
<td>90.1</td>
<td>47.5</td>
<td>33.4</td>
<td>41.0</td>
<td>37.5</td>
<td>51.1</td>
<td>70.5</td>
<td>60.8</td>
<td>71.0</td>
</tr>
</tbody>
</table>

Figure 1. Visualization results on the test set of ScanNet-V2. Raw point cloud, results of the baseline and ours are presented separately from top to bottom.

Evaluation on Semantic3D. We conduct the quantitative evaluations on Semantic3D (reduced-8) and list the per-class scores in Tab. 3. Mean Intersection-over-Union (mIoU) and Overall Accuracy (OA) of all classes are used as the standard metrics. We compared some full supervised methods published in recent years such as SnapNet [3],SEGCloud [13], ShellNet [19], KPConv [14], RandLANet [7], and PointGCR [9], RFCR [5]. At 1% setting, HybridCR achieves 76.8% and 94.9% in terms of both mIoU and OA, comparable to the fully-supervised methods. Compared with the fully supervised RandLA-Net, HybridCR is 0.6% lower in mIoU and 0.1% lower in OA, respectively. But, we achieve the best performance in the classes of “man-made” and “nature”. Therefore, the results show that HybridCR can generate to the sparse outdoor dataset.

Evaluation on SemanticKITTI. We conduct the quantitative evaluations on SemanticKITTI and list the per-class scores in Tab. 4. We compared some full supervised methods published in recent years, including PointNet [11], SqueezeSegV2 [15], DarkNet53Seg [2], RangeNet53++ [10] and RandLA-Net [7]. It can be found that HybridCR achieves the best performance among the fully-supervised setting comparison. At 1% setting, HybridCR reports 52.3% in mIoU, which are close to the performance of the fully-supervised methods. Compared with
1.4. Quantitative Results

Visualization on ScanNet-V2. In Fig. 1, we show visualization results on the test set of ScanNet-V2. Since there is no public ground truth, we show the raw point clouds at the top row and our segmentation results at the bottom row. It can be observed that HybridCR can achieve good segmentation results for most classes. At the 1% setting, the segmentation precision of small corners and boundaries e.g., “wall” and “door” area compared to PSD, is further improved.

Visualization on Semantic3D. Fig. 2 shows the visualization results on the test set of Semantic3D. Since there is no public ground truth, we show the raw point cloud at the top row and our segmentation results at the bottom row. In general, it can be seen that HybridCR achieves good qualitative segmentation results at 1% setting. Our method can also make more accurate predictions for some categories (e.g., “low-veg.” “buildings” and “man-made”) with a small number of points.

Visualization on SemanticKITTI. Fig. 3 shows more qualitative results of HybridCR on the validation split. It can be seen that our method achieves consistent segmentation results to ground-truth, especially in “road” and “car”, which are difficult to distinguish while critical on sparse outdoor scenes in the auto-driving application.

1.5. Dynamic vs. fixed augmentor of multiple runs.

In Tab. 5, we report 1pt and 1% results with mean and std.dev. (5 runs) on S3DIS Area-5, as well as dynamic and weakly augmented results. Our results show that HybridCR outperforms other methods in terms of mIoU and OA across all categories.
Table 4. Quantitative results of per class on SemanticKITTI [2]. (mIoU %)

<table>
<thead>
<tr>
<th>Set</th>
<th>Methods</th>
<th>road</th>
<th>sidewalk</th>
<th>parking</th>
<th>other-ground</th>
<th>building</th>
<th>car</th>
<th>truck</th>
<th>motorcycle</th>
<th>other-vehicle</th>
<th>vegetation</th>
<th>trunk</th>
<th>terrain</th>
<th>person</th>
<th>bicyclist</th>
<th>motorcyclist</th>
<th>fence</th>
<th>pole</th>
<th>traffic-sign</th>
</tr>
</thead>
</table>
| Fully     | PointNet [11]   | 14.6 | 61.6     | 35.7    | 15.8         | 8.4      | 41.4    | 46.3    | 0.1        | 1.3            | 0.3        | 0.8    | 31.0    | 4.6     | 17.6      | 0.2          | 0.0   | 12.9 | 2.4         | 3.7
|           | SqueezeSegV2 [15]| 39.7 | 88.6     | 67.6    | 45.8         | 17.7     | 73.7    | 81.8    | 13.4       | 18.5           | 17.9       | 14.0   | 71.8    | 35.8   | 60.2      | 20.1         | 25.1  | 41.1 | 20.2         | 36.3
|           | DarkNet53Seg [2] | 49.9 | 91.8     | 74.6    | 64.8         | 27.9     | 84.1    | 86.4    | 25.5       | 24.5           | 32.7       | 22.6   | 78.3    | 50.1   | 64.0      | 36.2         | 33.6  | 47.5 | 35.9         | 52.2
|           | RangeNet53++ [10]| 52.2 | 91.8     | 75.2    | 65.0         | 27.8     | 87.4    | 91.4    | 25.7       | 25.7           | 34.4       | 23.0   | 80.5    | 55.1   | 64.6      | 38.3         | 38.8  | 48.8 | 57.6         | 55.9
|           | RandLA-Net [7]  | 53.9 | 90.7     | 73.7    | 60.3         | 20.4     | 86.9    | 94.2    | 40.1       | 26.0           | 25.8       | 38.9   | 81.4    | 61.3    | 66.8      | 49.2         | 48.2  | 72.6 | 53.6         | 47.7
|           | HybridCR        | 54.0 | 90.5     | 73.9    | 69.3         | 21.2     | 88.3    | 93.9    | 42.7       | 22.8           | 31.6       | 36.8   | 81.7    | 61.7    | 66.1      | 50.2         | 45.5  | 57.4 | 49.5         | 49.0

| weakly    | HybridCR(1%)    | 52.3 | 89.4     | 72.9    | 61.5         | 20.6     | 85.8    | 64.5    | 50.1       | 45.8           | 3.9        | 55.2   | 41.8    | 48.2   | 64.5      | 69.3         | 50.1  | 45.8 | 41.8         | 48.2

Table 5. Dynamic vs. fixed augmentor on S3DIS Area-5 in 1pt and 1%. #1-#8 are the ablation settings in Tab. 3 of the main paper.

<table>
<thead>
<tr>
<th>Method</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
<th>#5</th>
<th>#6</th>
<th>#7</th>
<th>#8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1pt</td>
<td>48.2±(0.3)</td>
<td>50.7±(0.3)</td>
<td>49.8±(0.5)</td>
<td>50.2±(0.2)</td>
<td>51.1±(0.2)</td>
<td>50.8±(0.1)</td>
<td>51.0±(0.3)</td>
<td>51.5±(0.2)</td>
</tr>
<tr>
<td>Dynamic(1pt)</td>
<td>-</td>
<td>50.7±(0.3)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>50.8±(0.1)</td>
<td>51.1±(0.3)</td>
<td>51.5±(0.2)</td>
</tr>
<tr>
<td>Fix(1pt)</td>
<td>-</td>
<td>47.2±(0.4)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>47.7±(0.3)</td>
<td>48.0±(0.2)</td>
<td>48.3±(0.3)</td>
</tr>
<tr>
<td>1%</td>
<td>63.5±(0.1)</td>
<td>64.5±(0.3)</td>
<td>63.9±(0.4)</td>
<td>64.0±(0.2)</td>
<td>65.0±(0.3)</td>
<td>64.7±(0.4)</td>
<td>65.1±(0.2)</td>
<td>65.3±(0.3)</td>
</tr>
<tr>
<td>Dynamic(1%)</td>
<td>-</td>
<td>64.5±(0.3)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>64.7±(0.4)</td>
<td>65.1±(0.2)</td>
<td>65.3±(0.3)</td>
</tr>
<tr>
<td>Fix(1%)</td>
<td>-</td>
<td>59.8±(0.1)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>61.8±(0.2)</td>
<td>61.4±(0.3)</td>
<td>62.6±(0.1)</td>
</tr>
</tbody>
</table>

fixed augmentors. Note that #1-#8 are the ablation settings in Tab. 3 of the main paper. For 1pt and 1% setting, it can be found that the dynamic augmentor outperform the fixed one by 3.2% and 2.7% mIoU at the ablation setting #8, respectively.

References

Figure 3. Visualization results on the validation set of SemanticKITTI. Ground truth, results of the baseline and ours are presented separately from top to bottom.


