A. Performance with Different Backbones

In Tab. A1, alongside Cylinder3D [8], we also implement our architecture with popular backbone networks [2, 5] widely-used in 3D semantic segmentation.

B. Runtime Comparison

We conduct a runtime comparison between LiM3D and SORTAs on NVIDIA A100 GPU and Graphcore IPU-POD16 for Intelligence Processing Unit (IPU) acceleration. Summary inference times per item: LiM3D (ours, GPU) - 0.51s, Cylinder3D [8] (GPU) - 0.48s, Unal et al. [6] (GPU) - 0.53s, LiM3D+SDSC (ours, IPU) - 0.12s. SDSC could be slower on GPU due to its lower arithmetic intensity (ratio of compute to memory operation) [3, 7], but the high-bandwidth on-chip memory in IPU accelerator significantly improves the efficiency.

C. More Quantitative Results on Semi-supervised Segmentation

Besides {5%, 10%, 20%, 40%} labeled frames training, we also report our results with less than 5% label frames shown in Tab. A2. By applying our proposed architecture for semi-supervised and scribble-supervised 3D semantic segmentation, LiM3D and LiM3D+SDSC achieve higher than 80% relative performances (SS/FS) comparing with fully-supervised methods with less than only 1% labeled, i.e., 191 frames (Tab. A2).

D. More Qualitative Results

Figs. A1 and A2 show a higher-resolution version of qualitative results that our method has superior performance. Fig. A3 compares {5%, 10%, 20%, 40%} sampling splits of SemanticKITTI [1] using LiM3D (ours). Note that using our semi-supervised methodology, the results training with very few ground-truth labels (e.g., 5% and 10%) can achieve comparable performance to the training with a large number of labels (see Fig. A3, the 1-st and 2-nd rows v.s. 4-th row), with only subtle differences shown in the green and red circles. In Fig. A4, the magnification of regional details shows that our method can achieve better segmentation results than other methods, especially in the category of vegetation, fence, sidewalk, etc.

E. Parameters and Computation Costs Analysis on SDSC Sub-module

Given a Tensor $F \in \mathbb{R}^{H_F \times W_F \times L_F \times M}$, where $H_F$, $W_F$, $L_F$ and $M$ denote the radius, azimuth, height in the cylinder coordinate [8] and channels respectively. Applying convolution operation only for the active site of the sparse 3D point cloud, the computational cost (in FLOPs) of submanifold sparse convolution (SSC, [4]) is $a \times M \times N$ for the active site, where $M$ is the number of input channels as defined previously, and $N$ is the number of output channels. $a$ is the number of active inputs to the spatial location defined in [4]. The computational cost for the inactive site is 0.

Since our SDSC sub-module consists of a sparse depth-wise convolution (SDC) and a sparse pointwise convolution (SPC), the computational cost for SDSC is the sum of those two parts. SDC has a computational cost of

$$a \times M \times H_F \times W_F \times L_F.$$ \hspace{1cm} (A1)

SPC computes a linear combination of the SDC output via a $1 \times 1$ convolution, which has the computational cost of

$$M \times N \times H_F \times W_F \times L_F.$$ \hspace{1cm} (A2)

As a result, the computational cost of SDSC is the sum of Eqs. (A1) and (A2), i.e.,

$$a \times M \times H_F \times W_F \times L_F + M \times N \times H_F \times W_F \times L_F.$$ \hspace{1cm} (A3)

The ratio of computational cost of SDSC to SSC [4] for active site, i.e., cost(SDSC) : cost(SSC), is:

$$\frac{a \times M \times H_F \times W_F \times L_F + M \times N \times H_F \times W_F \times L_F}{a \times M \times N \times H_F \times W_F \times L_F} = \frac{1}{N} + \frac{1}{a} \approx \frac{1}{N} \ (a \gg N).$$ \hspace{1cm} (A4)

Similar to the computational cost analysis, the parameters of SDSC is also the sum of SDC and SPC. The ratio of model
Table A1. 3D semantic segmentation results of LiM3D (ours) evaluated on SemanticKITTI [1] and ScribbleKITTI [6] \textit{valid}-set with 10% labeled data, using different backbones. Alongside the per-class metrics, we show the relative performance of the semi-supervised approach against the fully supervised (SS/FS). S: with SDSC sub-module (✓) or without SDSC sub-module, \textit{i.e.}, with normal sparse convolution.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset</th>
<th>S</th>
<th>mIoU</th>
<th>SS/FF</th>
<th>car</th>
<th>bicycle</th>
<th>motorcycle</th>
<th>truck</th>
<th>other vehicle</th>
<th>person</th>
<th>bicyclist</th>
<th>motorcyclist</th>
<th>road</th>
<th>parking</th>
<th>sidewalk</th>
<th>other ground</th>
<th>building</th>
<th>fence</th>
<th>vegetation</th>
<th>trunk</th>
<th>terrain</th>
<th>pole</th>
<th>traffic</th>
<th>sign</th>
</tr>
</thead>
<tbody>
<tr>
<td>LiM3D (ours) + Cylinder3D [8]</td>
<td>Semantic</td>
<td>✓</td>
<td>62.2</td>
<td>89.5</td>
<td>95.5</td>
<td>47.6</td>
<td>65.2</td>
<td>60.7</td>
<td>42.4</td>
<td>75.2</td>
<td>84.3</td>
<td>0.0</td>
<td>94.2</td>
<td>42.4</td>
<td>80.7</td>
<td>5.4</td>
<td>91.0</td>
<td>61.1</td>
<td>86.6</td>
<td>65.7</td>
<td>70.7</td>
<td>64.0</td>
<td>49.2</td>
<td>65.8</td>
</tr>
<tr>
<td>LiM3D (ours) + MinkowskiNet [2]</td>
<td>Scribble</td>
<td>✓</td>
<td>60.4</td>
<td>86.9</td>
<td>94.6</td>
<td>44.3</td>
<td>47.1</td>
<td>70.2</td>
<td>29.5</td>
<td>68.7</td>
<td>80.8</td>
<td>0.0</td>
<td>93.6</td>
<td>38.4</td>
<td>79.6</td>
<td>0.1</td>
<td>92.0</td>
<td>58.7</td>
<td>82.2</td>
<td>66.1</td>
<td>75.6</td>
<td>65.5</td>
<td>58.9</td>
<td></td>
</tr>
<tr>
<td>LiM3D (ours) + SPVCNN [5]</td>
<td>Scribble</td>
<td>✓</td>
<td>56.2</td>
<td>80.9</td>
<td>93.8</td>
<td>42.7</td>
<td>37.6</td>
<td>68.4</td>
<td>33.7</td>
<td>54.2</td>
<td>63.5</td>
<td>0.0</td>
<td>91.9</td>
<td>39.9</td>
<td>76.7</td>
<td>0.1</td>
<td>88.9</td>
<td>61.9</td>
<td>86.9</td>
<td>67.3</td>
<td>74.4</td>
<td>59.1</td>
<td>27.7</td>
<td>68.0</td>
</tr>
</tbody>
</table>

Table A2. 3D semantic segmentation results of LiM3D (ours) evaluated on SemanticKITTI [1] and ScribbleKITTI [6] \textit{valid}-set with %1 and 2% labeled data. Alongside the per-class metrics, we show the relative performance of the semi-supervised approach against the fully supervised (SS/FS). S: with SDSC sub-module (✓) or without SDSC sub-module, \textit{i.e.}, with normal sparse convolution.

<table>
<thead>
<tr>
<th>% labeled</th>
<th>Dataset</th>
<th>S</th>
<th>mIoU</th>
<th>SS/FF</th>
<th>car</th>
<th>bicycle</th>
<th>motorcycle</th>
<th>truck</th>
<th>other vehicle</th>
<th>person</th>
<th>bicyclist</th>
<th>motorcyclist</th>
<th>road</th>
<th>parking</th>
<th>sidewalk</th>
<th>other ground</th>
<th>building</th>
<th>fence</th>
<th>vegetation</th>
<th>trunk</th>
<th>terrain</th>
<th>pole</th>
<th>traffic</th>
<th>sign</th>
</tr>
</thead>
<tbody>
<tr>
<td>LiM3D (ours) 2% / 383 frames</td>
<td>Semantic</td>
<td>✓</td>
<td>59.3</td>
<td>85.3</td>
<td>95.6</td>
<td>37.6</td>
<td>50.1</td>
<td>54.4</td>
<td>46.0</td>
<td>68.8</td>
<td>77.1</td>
<td>0.0</td>
<td>87.9</td>
<td>32.8</td>
<td>76.6</td>
<td>2.0</td>
<td>91.4</td>
<td>54.8</td>
<td>89.5</td>
<td>69.5</td>
<td>77.1</td>
<td>66.2</td>
<td>49.4</td>
<td></td>
</tr>
<tr>
<td>LiM3D (ours) 1% / 191 frames</td>
<td>Semantic</td>
<td>✓</td>
<td>58.7</td>
<td>84.5</td>
<td>94.8</td>
<td>37.1</td>
<td>55.0</td>
<td>56.2</td>
<td>45.2</td>
<td>66.1</td>
<td>75.4</td>
<td>0.0</td>
<td>87.0</td>
<td>32.5</td>
<td>75.9</td>
<td>2.1</td>
<td>89.1</td>
<td>49.7</td>
<td>89.3</td>
<td>68.0</td>
<td>76.0</td>
<td>65.8</td>
<td>45.9</td>
<td></td>
</tr>
<tr>
<td>LiM3D (ours) 1% / 191 frames</td>
<td>Scribble</td>
<td>✓</td>
<td>58.2</td>
<td>83.7</td>
<td>92.5</td>
<td>35.6</td>
<td>52.0</td>
<td>57.1</td>
<td>49.4</td>
<td>66.8</td>
<td>78.5</td>
<td>0.0</td>
<td>85.6</td>
<td>30.2</td>
<td>74.4</td>
<td>2.4</td>
<td>89.7</td>
<td>54.0</td>
<td>88.2</td>
<td>66.9</td>
<td>74.4</td>
<td>63.7</td>
<td>47.7</td>
<td></td>
</tr>
<tr>
<td>LiM3D (ours) 1% / 191 frames</td>
<td>Scribble</td>
<td>✓</td>
<td>56.8</td>
<td>81.7</td>
<td>93.1</td>
<td>34.9</td>
<td>47.0</td>
<td>50.9</td>
<td>43.8</td>
<td>64.1</td>
<td>75.6</td>
<td>0.0</td>
<td>85.2</td>
<td>29.5</td>
<td>73.9</td>
<td>2.0</td>
<td>88.8</td>
<td>49.5</td>
<td>88.2</td>
<td>66.9</td>
<td>74.4</td>
<td>64.4</td>
<td>47.4</td>
<td></td>
</tr>
</tbody>
</table>

parameters of SDSC to SSC [4] is:

$$D_K \times M \times H_F \times W_F \times L_F + M \times N \times H_F \times W_F \times L_F$$

$$D_K \times M \times N \times H_F \times W_F \times L_F$$

$$\frac{1}{N \cdot D_K} \approx \frac{1}{D_K} \quad (N \gg D_K),$$

(A5)

where $D_K$ is the dimension of convolution kernel $K$ of size $D_{K,1} \times D_{K,2} \times D_{K,3}, \text{ i.e.}, D_K = D_{K,1} \times D_{K,2} \times D_{K,3}.$

LiM3D+SDSC uses SDSC sub-module as the basic building block for constructing other convolution-based modules in Cylinder3D (e.g., residual block, upsample block, and downsample block). Take the residual block as an example, SDSC uses approximately 32x, 64x, $\cdots$, 512x less computation than SSC for active sites when $N = \{32, 64, \cdots, 512\}$ (Eq. (A4)). SDSC-based residual block with a kernel size of $1 \times 1 \times 1$ has 3x fewer parameters than the SSC-based residual block with the same kernel size, and 9x fewer parameters with $3 \times 1 \times 3$ kernel size (Eq. (A5)).
Figure A1. Comparing the 10% sampling split of SemanticKITTI [1] validation set with ground-truth (left), our approach (middle) and Unal et al. [6] (right) with areas of improvement highlighted in green, and areas of underperformance in red.
Figure A2. Comparing the 10% sampling split of SemanticKITTI [1] validation set with ground-truth (left), our approach (middle) and Unal et al. [6] (right) with areas of improvement highlighted in green, and areas of underperformance in red.
Figure A3. Comparing the 5%, 10%, 20%, 40% sampling split of SemanticKITTI [1] validation set with ground-truth (bottom) with areas of improvement highlighted in green, and areas of underperformance in red.
Figure A4. Magnification of regional details: comparing the 10% sampling split of SemanticKITTI [1] validation set with ground-truth (left), our approach (middle) and Unal et al. [6] (right) with areas of improvement highlighted in green, and areas of underperformance in red.
References


