TopNet: Transformer-Efficient Occupancy Prediction Network for Octree-Structured Point Cloud Geometry Compression

Supplementary Material

1. Detailed Model Complexity Analysis

In this section, we compare and analyze the computational cost between the standard Transformer and the linear projection layer (LPL) used in OctAttention and our TopNet. A standard Transformer block comprises the multi-head self-attention (MSA) module and the feed-forward network (FFN). For simplicity, we assume the Transformer consists of 3 layers. Given an input feature of size $N \times C$, the computational complexity measured in the number of **fl**oating-point **op**eration**s** (FLOPs) can be calculated as follows:

$$\Omega(\text{MSA}) = 6NC(C_k + C_v) + 3N^2(C_k + C_v) \tag{1}$$

$$\Omega(\text{FFN}) = 6NC^2R_1 \tag{2}$$

$$\Omega(\text{LPL}) = NC^2R_2 + 255NCR_2 \tag{3}$$

where R_1 and R_2 are the expansion ratios for FFN and LPL, respectively, while C_k and C_v represent the dimensions of the key and value. In OctAttention, specific settings are used $C = C_k = C_v$, $R_1 = 8$ and $R_2 = 1$, which simplifies the costs as follows:

$$\Omega(\text{MSA}) = 12NC^2 + 6N^2C \tag{4}$$

$$\Omega(\text{FFN}) = 48NC^2 \tag{5}$$

$$\Omega(\text{LPL}) = NC^2 + 255NC \tag{6}$$

$$\begin{aligned}\Omega(\text{OctAttention}) &= \Omega(\text{MSA}) + \Omega(\text{FFN}) + \Omega(\text{LPL}) \\ &= 61NC^2 + 6N^2C + 255NC\end{aligned} \tag{7}$$

For our proposed TopNet, the computational complexity is computed as follows:

$$\Omega(\text{LeCE}) = \frac{5}{4}kNC + \frac{3}{4}NC^2 \tag{8}$$

$$\Omega(\text{AL-SWA}) = 12NC^2 + 6N^2C + 3kNC \tag{9}$$

$$\Omega(\text{SG-CM}) = 36NC^2 + 12kNC \tag{10}$$

$$\Omega(\text{LNOP}) = 2kNC + 4NC^2 + 255NC \tag{11}$$

$$\Omega(\text{TopNet}) = \frac{73}{4}kNC + \frac{211}{4}NC^2 + 6N^2C + 255NC \tag{12}$$

To clearly illustrate the computational complexity, we set the depth-wise convolution kernel size to $k = 3$ and the number of input feature channels to $C = 256$. Substituting these values yields the following expressions for OctAttention and TopNet:

$$\Omega(\text{OctAttention}) = 4072896N + 1536N^2 \tag{13}$$

$$\Omega(\text{TopNet}) = 3537280N + 1536N^2 \tag{14}$$

Compared to OctAttention, the proposed TopNet demonstrates significantly reduced computational costs, primarily due to its streamlined architectural design. This reduction in complexity enhances its efficiency, making it more practical for deployment in real-world applications, particularly those with limited computational resources.

2. Effect of LeCE Module

In this section, we examine the contributions of the local perception unit (LPU) and inception-residual network (IRN) blocks within the LeCE module. The results are summarized in Table 1, and demonstrate the effectiveness of these components in improving compression performance.

- **Baseline Model without Positional Encoding (w/o PE).** Adding absolute positional encoding (APE) resulted in slight performance drops, with BPP increasing by 0.71% for the 8iVFB dataset and 0.84% for the MVUB dataset. This indicates that the original APE has minimal impact on the model's performance.
- **LPU Block Only.** Incorporating only the LPU block led to BPP reductions of 0.53% for 8iVFB and 0.56% for MVUB, emphasizing the role of the LPU in enhancing the performance of LeCE.
- **IRN Block Only.** Adding only the IRN block further improved performance, with BPP reductions of 0.88% for 8iVFB and 0.98% for MVUB, highlighting the efficiency of the IRN in feature extraction.
- **LPU and IRN Combined.** The combination of both blocks yielded the most significant improvements, with BPP reductions of 1.24% for 8iVFB and 1.39% for MVUB. This demonstrates the synergy between the LPU and IRN in capturing local spatial structures and inter-channel dependencies.

3. Effect of AL-SWA Module

This section investigates the contribution of various components within the AL-SWA module, including depth-wise convolution (DWConv), query embedding (QE), scaled dot product attention (SDPA), scaled cosine attention (SCA), and length-scaled cosine attention (LSCA). As shown in Table 2, the use of DWConv alone results in BPP reductions of 2.23% for 8iVFB and 1.63% for MVUB compared to the baseline model (MSA). When DWConv is combined with QE, it achieves further BPP reductions of 3.08% for 8iVFB and 2.59% for MVUB. The introduction of SCA, an improvement over SDPA, results in even greater reduc-

tions of 3.60% for 8iVFB and 3.00% for MVUB. Furthermore, the introduction of LSCA, which improves upon both SCA and SDPA, leads to even more significant reductions of 4.45% for 8iVFB and 3.68% for MVUB, underscoring the substantial performance gains offered by LSCA over the traditional scaled dot product and scaled cosine attention mechanisms.These results suggest that the integration of efficient local feature extraction through DWConv, enhanced attention mechanisms with QE and LSCA, and the synergistic effects of these components significantly boosts the performance of the module.

Analysis. the improvements observed can be attributed to the effective combination of local feature extraction using DWConv and advanced attention mechanisms provided by QE and LSCA. This combination enhances efficiency and ensures that the model can better capture both local and global dependencies in the data.

4. Effect of SG-CM Module

This section investigates the impact of integrating spatial convolution operations within the SG-CM on compression performance. The empirical findings are presented in Table 3. After adding the DWConv operation, the SG-CM achieved BPP reductions of 0.53% and 0.56% for the 8iVFB and MVUB datasets, compared to the initial baseline model FFN. Subsequently, introducing the split (ST) operation, which decomposes along the channel dimension and employs a gating mechanism to effectively reduce the number of model parameters, led to BPP reductions of 1.23% and 1.25%. Finally, adding the shortcut (SC) operation further decreased the BPP by 1.59% and 1.53% for the 8iVFB and MVUB datasets, respectively.

Analysis. These results demonstrate the effectiveness of the SG-CM in improving compression performance. The observed performance gains are attributed to the efficient feature extraction and parameter reduction enabled by the combination of DWConv, ST, and SC operations. These strategies highlight the value of integrating diverse convolutional techniques to enhance compression efficiency.

5. Effect of LNOP Module

This section investigates the ablation experiments on LNOP to assess the impact on compression performance. The results are presented in Table 4. When only the Sum operation was used, BPP decreased by 0.18% and 0.28% for the 8iVFB and MVUB datasets, respectively. Adopting the Star operation in the LNOP led to BPP reductions of 0.53% and 0.56%, indicating that the Star operation performs better in nonlinear fitting and prediction. Finally, adding the SC operation further improved the model, resulting in BPP reductions of 0.71% and 0.70% for the 8iVFB and MVUB datasets, respectively.

Analysis. These results highlight the effectiveness of the LNOP module in improving compression performance. Both the Star and SC operations contribute to these improvements, with the performance gains attributed to enhanced nonlinear fitting capabilities and the efficient utilization of residual connections.

6. Additional Qualitative Results of Reconstruction Quality for Sparse LiDAR Point Clouds

The qualitative results comparing compression distortions at lower BPP between our method and baseline method at similar bitrates are shown in Fig. 1. Our method achieves higher D1 PSNR values and lower CD metrics, as shown in Fig. 1. This indicates that the reconstructed point clouds retain greater structural fidelity, with fewer artifacts, and closely approximate the original input in terms of the number of reconstructed points. Higher D1 PSNR values indicate that our method preserves structural fidelity more effectively, producing reconstructed point clouds with fewer artifacts. Additionally, the lower CD metric demonstrates that the geometric shapes of the reconstructed point clouds more closely resemble their uncompressed counterparts, further highlighting the effectiveness of our method in sparse point cloud scenarios.

Module	LPU	IRN	BPP (↓) on 8iVFB	BPP (↓) on MVUB	Param.
w/o PE			0.565	0.717	3.318M
APE			0.569 (+0.71%)	0.723 (+0.84%)	3.318M
LeCE	✓		0.562 (−0.53%)	0.713 (−0.56%)	3.319M
LeCE		✓	0.560 (−0.88%)	0.710 (−0.98%)	3.368M
LeCE	✓	✓	**0.558 (−1.24%)**	**0.707 (−1.39%)**	**3.369M**

Table 1. Results of ablation experiments on the 8iVFB and MVUB datasets for LeCE. Note that LPU denotes the local perception unit, and IRN stands for the inception-residual network.

Module	DWConv	QE	SDPA	SCA	LSCA	BPP (↓) on 8iVFB	BPP (↓) on MVUB	Param.
MSA			✓			0.584	0.734	3.359M
AL-SWA	✓		✓			0.571 (−2.23%)	0.722 (−1.63%)	3.368M
AL-SWA	✓	✓	✓			0.566 (−3.08%)	0.715 (−2.59%)	3.369M
AL-SWA	✓	✓		✓		0.563 (−3.60%)	0.712 (−3.00%)	3.368M
AL-SWA	✓	✓			✓	**0.558 (−4.45%)**	**0.707 (−3.68%)**	**3.369M**

Table 2. Results of ablation experiments on the 8iVFB and MVUB datasets for AL-SWA. Note that DWConv denotes the depth-wise convolution operation, QE stands for the query embedding strategy, SDPA denotes the scaled dot product attention, SCA denotes the scaled cosine attention, and LSCA represents the length-scaled cosine attention.

Module	DWConv	ST	SC	BPP (↓) on 8iVFB	BPP (↓) on MVUB	Param.
FFN				0.567	0.718	4.153M
SG-CM	✓			0.564 (−0.53%)	0.714 (−0.56%)	4.164M
SG-CM	✓	✓		0.560 (−1.23%)	0.709 (−1.25%)	3.369M
SG-CM	✓	✓	✓	**0.558 (−1.59%)**	**0.707 (−1.53%)**	**3.369M**

Table 3. Results of ablation experiments on the 8iVFB and MVUB datasets for SG-CM. Note that DWC denotes the depth-wise convolution operation, ST stands for the split operation, and SC represents the shortcut operation.

Module	Sum	Star	SC	BPP (↓) on 8iVFB	BPP (↓) on MVUB	Param.
LPL				0.562	0.712	3.169M
LNOP	✓			0.561 (−0.18%)	0.710 (−0.28%)	3.369M
LNOP		✓		0.559 (−0.53%)	0.708 (−0.56%)	3.369M
LNOP		✓	✓	**0.558 (−0.71%)**	**0.707 (−0.70%)**	**3.369M**

Table 4. Results of ablation experiments on the 8iVFB and MVUB datasets for LNOP. Note that Sum denotes the element-wise addition operation, Star stands for the element-wise product operation, and SC represents the shortcut operation.

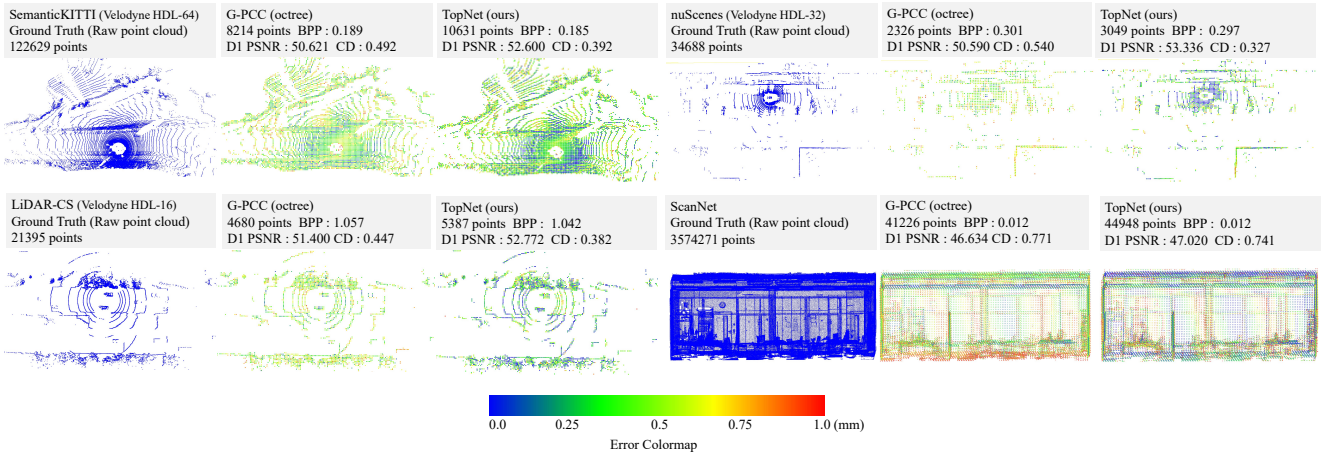

Figure 1. Visualization of reconstructed point clouds at lower BPP for G-PCC (octree), OctAttention, and our TopNet across the SemanticKITTI, nuScenes, LiDAR-CS, and ScanNet datasets.