# –Supplementary Material–

# ConDaFormer: Disassembled Transformer with Local Structure Enhancement for 3D Point Cloud Understanding

**Anonymous Author(s)**
Affiliation
Address
email

## S1 Ablation of window size

As stated in the main paper, our proposed disassembled window attention offers a notable advantage over the vanilla 3D cubic window attention by significantly reducing computational effort, enabling the potential enlargement of the receptive field through an increase in the window size. However, we also acknowledge a limitation of our ConDaFormer, which is the degradation in performance when utilizing a larger attention window. In this section, we present detailed ablation results to further investigate this issue. To assess the impact of window size on the performance of ConDaFormer, we conduct experiments using varying window sizes: $\{0.32m, 0.48m, 0.64m\}$, and present the corresponding results in Table S1. It is worth noting that as the window size expands, the training loss ($\mathrm{Loss_t}$) gradually decreases, and the performance on the training set ($\mathrm{mIoU_t}$) steadily improves. However, contrary to expectations, the performance on the validation set experiences a decline, indicating the occurrence of over-fitting. This phenomenon aligns with the observation made in LargeKernel [1] when increasing the convolutional kernel size. The introduction of a larger attention window incorporates additional positional embeddings, potentially resulting in optimization difficulties and leading to over-fitting. To address this issue, future research can explore techniques such as self-supervised or supervised pre-training on large-scale datasets. These approaches have shown promise in mitigating over-fitting and improving generalization performance. By leveraging such techniques, it is possible to enhance the robustness of ConDaFormer and enable the utilization of larger attention windows without suffering from performance degradation.

## S2 Ablation of position encoding

To enhance the modeling of crucial position information necessary for self-attention learning, we employ the contextual relative position encoding (cRPE) scheme introduced by Stratified Transformer [2]. In this context, we compare the performance of cRPE with two alternative position encoding schemes: Swin [3], wherein the learned relative position bias is directly added to the similarity between query and key, and PTv2 [4], which generates the position bias through an MLP that takes the relative position as input and subsequently adds it to the similarity between query and key. As shown in Table S2, cRPE outperforms the other schemes in two out of three metrics, indicating the significance of contextual features in effectively capturing fine-grained position information.

Table S1. Ablation of window size.

| Window | mIoU | mAcc | OA | $\mathrm{Loss}_t$ | $\mathrm{mIoU}_t$ |
|---|---|---|---|---|---|
| 0.32m | **72.6** | **78.4** | **91.6** | 0.52 | 95.8 |
| 0.48m | 72.1 | 78.3 | 91.5 | 0.47 | 96.0 |
| 0.64m | 71.6 | 78.4 | 91.4 | **0.45** | **96.1** |

Table S2. Ablation of position encoding.

| Method | mIoU | mAcc | OA |
|---|---|---|---|
| cRPE | **70.7** | **76.9** | 90.6 |
| Swin | 69.6 | 76.1 | 90.6 |
| PTv2 | 70.1 | 76.4 | **91.2** |

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

## References

[1] Yukang Chen, Jianhui Liu, Xiaojuan Qi, Xiangyu Zhang, Jian Sun, and Jiaya Jia. Scaling up kernels in 3d cnns. *Adv. Neural Inform. Process. Syst.*, 2022. 1

[2] Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3d point cloud segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8500–8509, 2022. 1

[3] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Int. Conf. Comput. Vis.*, 2021. 1

[4] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. In *Adv. Neural Inform. Process. Syst.*, 2022. 1