GeoAuxNet: Towards Universal 3D Representation Learning for Multi-sensor Point Clouds

Supplementary Material

1. Additional Experiments

1.1. Experiment Settings

Training Methodology. The comprehensive configuration for the joint pre-training and subsequent fine-tuning phases is shown in Table 1. The GeoAuxNet model is subjected to joint pre-training utilizing three distinct datasets: S3DIS [1] and ScanNet [4] derived from RGB-D camera, and SemanticKITTI [2] obtained via LiDAR. To account for variations in dataset scale, we established a sampling ratio of 2:2:5 across these datasets. Subsequently, the pre-trained GeoAuxNet model undergoes fine-tuning on each dataset independently, employing a reduced learning rate. The total number of training iterations is equal to the sum of the best performance necessary iteration numbers for all three datasets.

Table 1. Detailed training settings of semantic segmentation experiments.

Config	Pre-training
optimizer	SGD
scheduler	OneCycleLR
learning rate	0.05
weight decay	10^{-4}
momentum	0.9
batch size	24
epoch	100

Network Architectures. The detailed information of our backbone and point network is outlined in Table 2. The point network only contains 1.1M parameters which can be ignored compared with the voxel backbone. But it improves the performance by about 6% in mIoU on three datasets, as illustrated in Table 3.

Table 2. Details of the network architectures in GeoAuxNet.

Config	Voxel backbone	Point network
embedding channels	32	32
encoder layers	[2, 3, 4, 6]	[2, 2, 2, 2]
encoder channels	[32, 64, 128, 256]	[32, 64, 128, 256]
decoder layers	[2, 2, 2, 2]	[2, 2, 2, 2]
decoder channels	[256, 128, 96, 96]	[256, 128, 96, 96]

Table 3. Semantic segmentation results on three benchmarks. We train the voxel backbone and GeoAuxNet on the joint training data of three datasets. We report the mIoU (%) on Area 5 of S3DIS and validation sets of ScanNet and SemanticKITTI.

Methods	S3DIS	ScanNet	SemanticKITTI
Voxel backbone	63.4	64.7	57.9
GeoAuxNet	$69.5_{(+6.1)}$	$71.3_{(+6.6)}$	$63.8_{(+5.9)}$

1.2. Additional Results

We further conduct experiments with different training datasets. As shown in Table 4, the observed improvements are consistent on different selections of datasets.

Table 4. Semantic segmentation results on different selections of training datasets. The yellow columns stand for the results of three methods trained on S3DIS and nuScenes collectively, while the blue columns are the results of training on ScanNet and nuScenes. We report the mIoU (%) on Area 5 of S3DIS and validation sets of ScanNet and nuScenes.

Methods	S3DIS	nuScenes	ScanNet	nuScenes
SPVCNN [6]	44.1	56.9	46.8	58.4
PPT [8]	63.9	65.4	65.3	65.8
GeoAuxNet	**68.4**	**70.8**	**69.8**	**71.6**

1.3. Efficiency Analysis

Auxiliary learning aims to improve the model performance on the primary task by exploiting beneficial information from auxiliary tasks, while auxiliary tasks can be removed during inference. Our idea is to design a voxel network for the primary task to maintain its efficiency, and a point network for the auxiliary task so that it provides geometric information and is removed during inference. While kNN limits the efficiency, during inference we only preserve the Geometry Pool and Geo-to-Occ Auxiliary modules without the point network and kNN. As shown in Table 5, we validate the efficiency of different models on various datasets.

2. Additional Visualization

We provide more visualizations for MinkowskiNet [3], SPVCNN [6], PPT [8] and GeoAuxNet in Figure 1. The limitation of learning sensor-specific information in MinkowskiNet leads to unsatisfactory performance on the

Table 5. The inference time and throughput on data with a single NVIDIA A6000 GPU.

Method	Params.	Inference Time (ms) ↓			Throughput (ins./sec.) ↑		
		ScanNet	SemanticKITTI	nuScenes	ScanNet	SemanticKITTI	nuScenes
PointNet++ [5]	1.0M	1987	2013	2195	96	880	300
PT [9]	7.8M	5779	5814	6298	32	268	95
PTv2 [7]	3.9M	24275	24834	27695	10	96	31
MinkowskiNet [3]	60.9M	237	245	275	728	31	2490
SPVCNN [6]	61.0M	246	244	284	550	2490	1922
PPT [8]	63.0M	402	407	471	210	1922	692
GeoAuxNet	64.7M	267	269	303	462	692	1589

joint training benchmark. PPT only utilize voxel representations, while the point branch in SPVCNN does not provide fine-grained geometric features. The introduction of elaborate geometric information in GeoAuxNet preserves better detailed structures for point clouds from various sensors.

3. Additional Discussion and Future Works

Benefiting from extensive training data, universal models have achieved remarkable performance in natural language processing and 2D vision. PPT [8] studies cross-dataset learning in 3D vision which focuses on point clouds in different datasets from the same sensor. However, the domain gap between point clouds from different sensors still limits the university of 3D networks, which hampers the fusion and utilization of data from diverse sensors in practice. We propose GeoAuxNet to address this issue in an efficient way towards universal 3D representation learning. However, high quality 3D data is limited compared with the large corpus and numerous images. Therefore, our future researches will be undertaken to leverage text and 2D information for 3D universal models. Besides, more work will need to be done for the generation of scene-level 3D data.

References

[1] Iro Armeni, Ozan Sener, Amir Roshan Zamir, Helen Jiang, Ioannis K. Brilakis, Martin Fischer, and Silvio Savarese. 3D Semantic Parsing of Large-Scale Indoor Spaces. *CVPR*, pages 1534–1543, 2016. 1, 3

[2] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, C. Stachniss, and Juergen Gall. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. *ICCV*, pages 9296–9306, 2019. 1, 3

[3] Christopher Bongsoo Choy, JunYoung Gwak, and Silvio Savarese. 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. *CVPR*, pages 3070–3079, 2019. 1, 2

[4] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. *CVPR*, pages 2432–2443, 2017. 1, 3

[5] C. Qi, L. Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *NIPS*, 2017. 2

[6] Haotian* Tang, Zhijian* Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching Efficient 3D Architectures with Sparse Point-Voxel Convolution. In *ECCV*, 2020. 1, 2

[7] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point Transformer V2: Grouped Vector Attention and Partition-based Pooling. In *NIPS*, 2022. 2

[8] Xiaoyang Wu, Zhuotao Tian, Xin Wen, Bohao Peng, Xihui Liu, Kaicheng Yu, and Hengshuang Zhao. Towards Large-scale 3D Representation Learning with Multi-dataset Point Prompt Training. *ArXiv*, abs/2308.09718, 2023. 1, 2

[9] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, and Vladlen Koltun. Point Transformer. In *ICCV*, 2021. 2

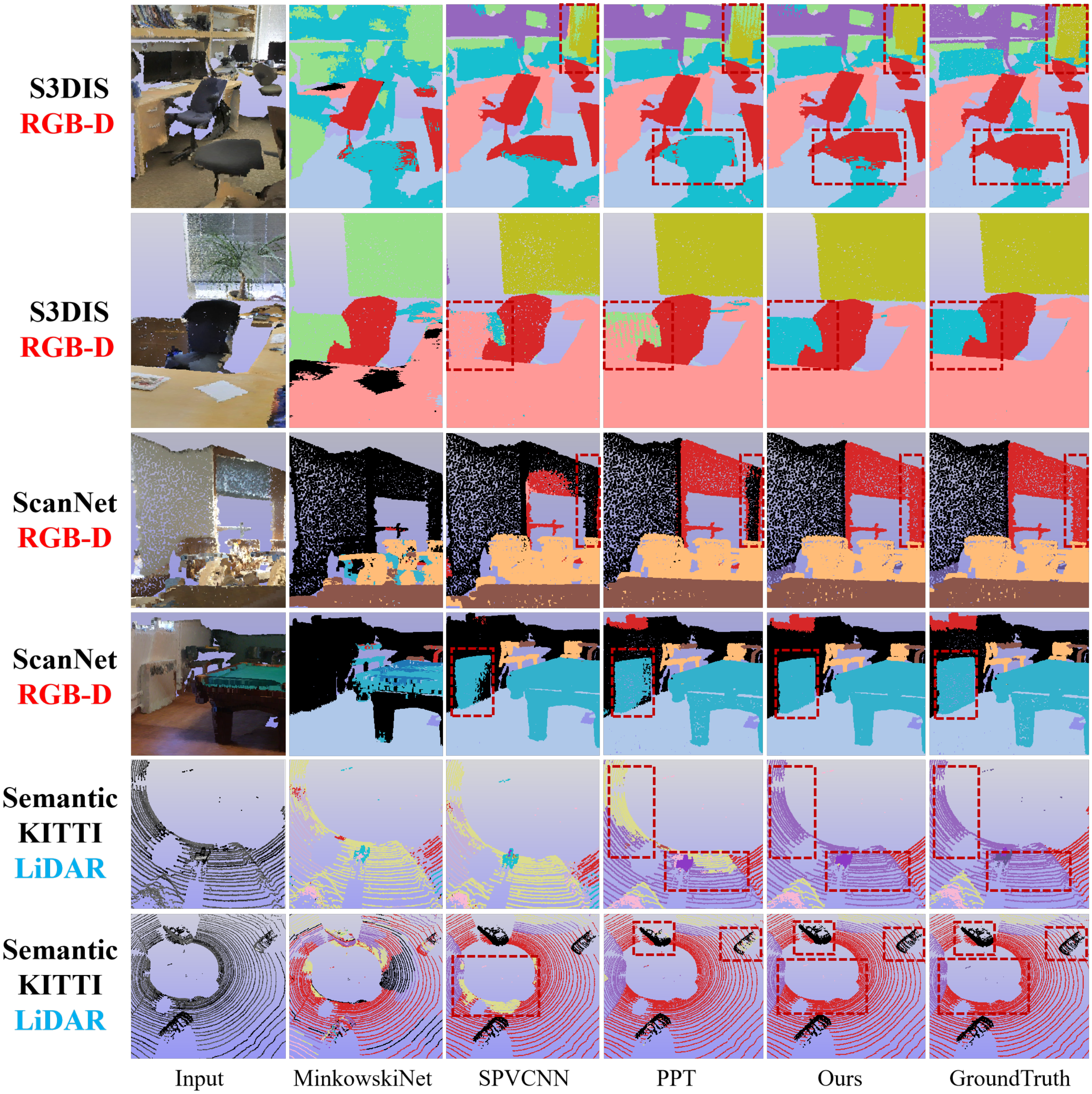

Figure 1. Addtional semantic segmentation results on S3DIS [1] and ScanNet [4] from RGB-D cameras and SemanticKITTI [2] from LiDAR. All methods are trained collectively on three datasets.