Supplementary Material of Point-to-Voxel Knowledge Distillation for LiDAR Semantic Segmentation

Yuenan Hou[1], Xinge Zhu[2], Yuexin Ma[3], Chen Change Loy[4], and Yikang Li[1]
[1]Shanghai AI Laboratory [2]The Chinese University of Hong Kong
[3]ShanghaiTech University [4]S-Lab, Nanyang Technological University
[1]{houyuenan, liyikang}@pjlab.org.cn, [2]zhuxinge123@gmail.com,
[3]mayuexin@shanghaitech.edu.cn, [4]ccloy@ntu.edu.sg

Table 1. Performance of different algorithms on distilling SPVNAS and MinkowskiNet on SemanticKITTI validation set.

Algorithm	mIoU	MACs (G)
SPVNAS [2]	**63.8**	118.6
SPVNAS_0.5×	60.4	29.7
SPVNAS_0.5× + KD	60.6	
SPVNAS_0.5× + CD	60.9	
SPVNAS_0.5× + IFV	60.8	
SPVNAS_0.5× + SKD	61.2	
SPVNAS_0.5× + KA	60.7	
SPVNAS_0.5× + PVD	**63.8**	
MinkowskiNet [1]	61.9	114.0
MinkowskiNet_0.5×	58.9	28.5
MinkowskiNet_0.5× + KD	59.2	
MinkowskiNet_0.5× + CD	59.6	
MinkowskiNet_0.5× + IFV	59.1	
MinkowskiNet_0.5× + SKD	59.4	
MinkowskiNet_0.5× + KA	59.2	
MinkowskiNet_0.5× + PVD	61.8	

1. Quantitative results

We provide the complete quantitative results of different algorithms on SPVNAS [2] and MinkowskiNet [1] in Table 1. Apparently, PVD consistently outperforms previous distillation algorithms by a large margin. For instance, on SPVNAS, PVD can bring 2.6 more points than the SKD method in mIoU. For both models, PVD can almost mitigate the performance gap between the original network and the pruned model. The encouraging results on SPVNAS and MinkowskiNet convincingly demonstrates the good scalability of PVD.

2. Ablation studies

Loss coefficients. By comparing each row with the last row in Table 2, we have the following observations: 1) the loss coefficient of the inter-voxel affinity distillation should be larger than other distillation losses to yield the best distillation effect (row 1 and 2). 2) Exchanging the loss coefficients of the point-based distillation loss and voxel-based distillation loss deteriorates the performance, which means the voxel-based distillation loss guides the point-based loss and is more important (row 5). 3) Slightly increasing the loss coefficients will not significantly affect the overall performance, which demonstrates the robustness of PVD (row 3 and 4).

Table 2. Performance of using different loss coefficients for PVD.

α_1	α_2	β_1	β_2	mIoU
0.1	0.1	0.1	0.1	65.2
0.25	0.25	0.25	0.25	65.5
0.15	0.15	0.15	0.25	66.2
0.1	0.2	0.2	0.25	66.3
0.15	0.1	0.25	0.15	65.4
0.1	0.15	0.15	0.25	66.4

Table 3. Influence of each component on the final performance.

$\mathcal{L}_{out\_p}$	$\mathcal{L}_{out\_v}$	$\mathcal{L}_{aff\_p}$	$\mathcal{L}_{aff\_v}$	mIoU
				63.1
√				63.4
	√			63.7
		√		63.6
			√	64.5

Performance sensitiveness to f_{class}. We conduct experiments on examining the effect of f_{class}. We rewrite the f_{class} to be $f_{class} = \alpha \exp(\beta N_{minor}) + 1$. Then, we randomly choose α from $\{3, 4, 5, 6\}$ and β from { -1, -2, -3}, and compare the performance of different combinations. Experimental results reveal that the final performance of PVD on Cylinder3D_0.5× ranges from 66.2 to 66.4. The small fluctuations in distillation performance indicates that PVD is not very sensitive to f_{class}.

The influence of each component of PVD. Detailed performance of each component is summarized in Table 3. The voxel-based loss term indeed has larger impacts on the final performance. One potential reason is that the voxel representation provides richer structural information of the environment as it aggregates the information of all points within a voxel.

Broader impact of PVD. We apply PVD to SemanticKITTI multi-scan segmentation tasks and observe **4.2**% performance improvement on the Cylinder3D_0.5× backbone. The resulting model ranks 3rd on the SemanticKITTI multi-scan competition [1].

Performance w.r.t the distances of objects. We repartition SemanticKITTI according to the position of the cars. Cars in the training set are relatively close to the origin ($\leq$ 20 m) while cars in the validation set are relatively far away from the origin ($>$ 20 m). We apply PVD to distill the Cylinder3D model on the newly divided dataset. Experimental results show that PVD can still bring **4.7**% to the Cylinder3D_0.5× model on cars (91.3% v.s. 95.6%).

3. Elaborated implementation details

For the baseline knowledge distillation approaches, the value of each loss coefficient is provided in Table 4. Since training a single model from scratch may take one more week, we resort to loading the pre-trained weights to accelerate the training process. In this condition, the overall training duration will be shortened to three days. Note that all methods adopt this strategy to ensure fair comparison. The latency is recorded using a single GPU (NVIDIA Tesla PG503-216 GV100) and the final value of latency is obtained after averaging the latency of 100 samples. The training protocol for SPVNAS and MinkowskiNet is exactly the same as their open-sourced codes[2]. Finetuning denotes retraining the trained model for 10 more epochs with the learning rate being 2e-4.

MACs calculation: Since sparse convolution merely operates on the non-zero positions and different input point cloud sequence has different non-zero patterns, we first estimate the average kernel map size following [2] for each layer and then use the following equation to compute the FLOPs of each layer: $FLOPs = N \times K_s \times C_{in} \times C_{out}$, where K_s is the size of kernel map, N is the number of points, C_{in} is the number of input channels, C_{out} is the number of output channels.

4. Qualitative results

We provide more visual comparisons of PVD with previous distillation algorithms in Fig. 1. Compared with the rival SKD approach, our PVD can significantly improve the prediction of the student model. The prediction errors of PVD on those minority classes, *e.g.*, person and bicycle, are much smaller than those of SKD. Besides, on objects that are faraway from the origin, *e.g.*, the car highlighted by the green rectangle, PVD also yields more accurate predictions than SKD. And PVD has lower inter-class similarity and higher intra-class similarity than SKD. The aforementioned results explicitly demonstrate the efficacy of PVD in distilling structural knowledge from the teacher model to the student.

References

[1] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 1

[2] 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 *European Conference on Computer Vision*, pages 685–702. Springer, 2020. 1, 2

[1] https://competitions.codalab.org/competitions/20331#results (multi-scan competition) till 2021-12-1 00:00 Pacific Time, and our method is termed PV-KD

[2] https://github.com/mit-han-lab/spvnas

Table 4. Loss coefficients of different distillation methods.

Model	KD	IFV	SKD		CD		KA		PVD			
			λ_{pi}	λ_{pa}	λ_{fea}	λ_{score}	λ_{ada}	λ_{aff}	α_1	α_2	β_1	β_2
Cylinder3D	0.2	0.2	0.15	0.2	0.1	0.3	0.2	0.15	0.1	0.15	0.15	0.25
SPVNAS	0.1	0.3	0.15	0.15	0.05	0.1	0.2	0.1	0.1	0.1	0.1	0.15
MinkowskiNet	0.2	0.2	0.1	0.2	0.05	0.1	0.15	0.2	0.1	0.1	0.15	0.2

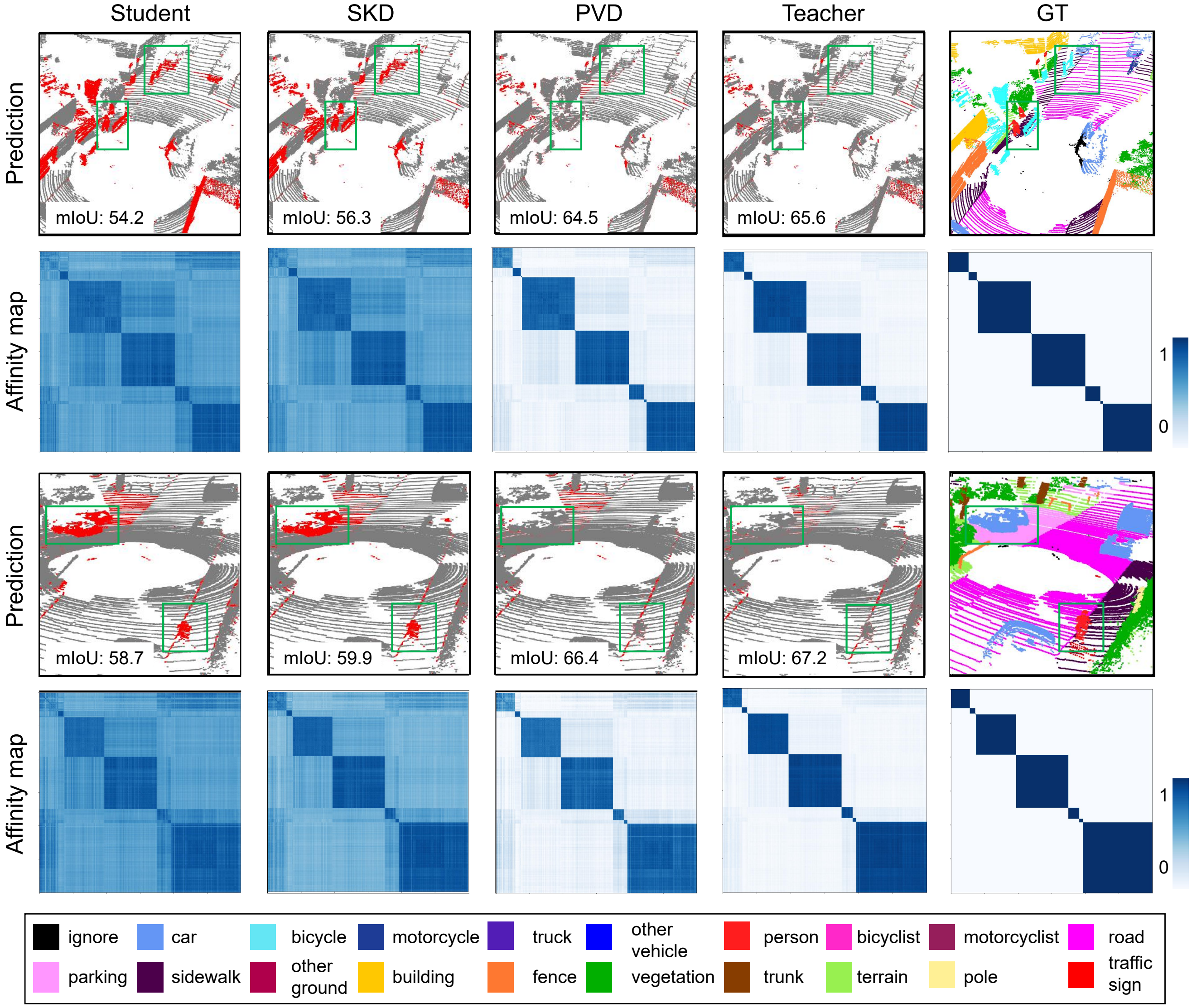

Figure 1. Visual comparison of different methods on the SemanticKITTI validation set. Here, the ground-truth for the inter-voxel affinity map is the ideal map where the intra-class similarity score is 1 and inter-class similarity score is 0.