2DPASS: 2D Priors Assisted Semantic Segmentation for LiDAR Point Clouds
Supplementary Material

A Overview

In this supplementary material, we first provide more implementation details in Section B. Besides, we provide more experiments and analysis in Section C. Finally, we provide concrete results, including results on validation sets and public online benchmarks.

B Training and Inference Details

This section provides training and inference details of the proposed 2DPASS. For the 3D input, we utilize the widely used data augmentation strategy for semantic segmentation, including global scaling with a random scaling factor sampled from [0.95, 1.05], and global rotation around the Z axis with a random angle. For the 2D input, we employ horizontal flipping and color jitter. Each 2D image is cropped to the size 480 $\times$ 320 (width $\times$ height) for faster training. The 2DPASS is trained in an end-to-end manner from scratch with the SGD optimizer. For the SemanticKITTI dataset, our model was trained with batch size 8 and learning rate 0.24 for 64 epochs, where the cosine annealing learning rate strategy is adopted for the learning rate decay. Moreover, we adopt instance-level augmentation [17] for a better performance on 'motorcyclist' category. As for the NuScenes dataset, we trained the model with batch size 16 for 80 epochs since the number of points per scene in NuScenes is generally smaller. During the inference, following [1,2], we apply the test-time augmentation, *i.e.*, rotating the input scene with eight angles around the Z axis and averaging the prediction scores. All experiments are conducted using a single Nvidia Tesla V100 GPU.

C Additional Experiments

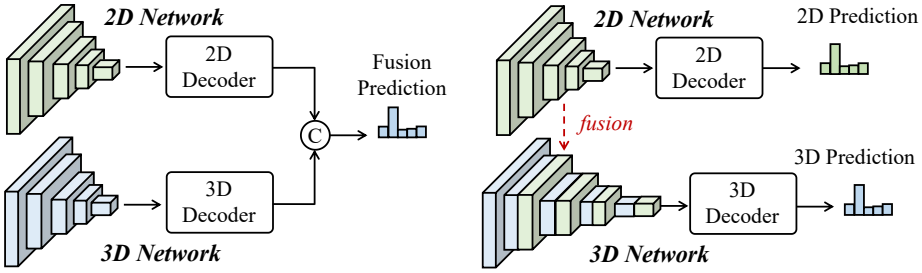
C.1 Comparing with Multi-Sensor Architecture

To further demonstrate the advantages of our 2DPASS upon multi-sensor methods, we set several multi-sensor baselines as well and compare against them in this section.

– **PointPainting**: We follow the setup of previous work [7], which exploits the segmentation logits of images and projects them to the LiDAR space by

Table 1. Comparison with different multi-sensor manners.

Method	mIoU (%)	Speed (ms)
PointPainting [7]-FCN-ResNet34 [11]	76.54	2330
PointPainting [7]-DeepLabV3 [12]	76.56	3347
Multi-branch Baseline	77.25	2353
Multi-branch with Interaction	**79.12**	2374
Basline (Tiny)	76.04	40
2DPASS (Tiny)	78.87	40



(a) Multi-branch Baseline *(b) Multi-branch with Interaction*

Fig. 1. The illustration of multi-sensor methods.

bird's-eye projection [8] or spherical projection [9]. Here, we use several pre-trained backbones, *i.e.*, FCN [10] with ResNet34 [11] and DeepLab_v3 [12], to achieve the 2D semantic segmentation logits. After that, we use outputs of 2D backbones as the inputs of our 3D network.

- **Multi-branch Baseline**: As shown in Fig. 1 (a), we design an ensemble architecture through concatenating the output logits from the two modalities.
- **Multi-branch with Interaction**: Instead of only concatenating the predictions, we also concatenate the 2D features from each layer into the corresponding layers in the 3D network, as illustrated in Fig. 1 (b).
- **2DPASS (Tiny)**: Since above multi-sensor manners are trained with the entire 2D image as input, they are time-consuming and GPU memory cost expensive. So we set all of hidden dimensions as 64 in the 3D network due to GPU memory limitation. This design is different from our manuscript with hidden dimensions 128 due to our light memory cost.

The experiment results are shown in Tab. 1, where we illustrate the results on NuScenes validation set and inference time (speeds), respectively. As shown in Tab. 1, using naive combination such as PointPainting [7] and concatenation (*i.e.*, Multi-branch Baseline) of prediction cannot improve the segmentation results obviously while introducing huge computational burden (*i.e.*, there are six 1600×900 camera images corresponding to each point cloud). Exploiting feature combination in each scale can slightly improve the performance, but leads to much slower network compared with the pure 3D network. On the contrary,

Table 2. Semantic segmentation results on the NuScenes valid set.

Method	Input	**mIoU**	barrier	bicycle	bus	car	construction	motorcycle	pedestrian	traffic cone	trailer	truck	driveable	other flat	sidewalk	terrain	manmade	vegetation
$(AF)^2$-S3Net [13]	L	62.2	60.3	12.6	82.3	80.0	20.1	62.0	59.0	49.0	42.2	67.4	94.2	68.0	64.1	68.6	82.9	82.4
RangeNet++ [9]	L	65.5	66.0	21.3	77.2	80.9	30.2	66.8	69.6	52.1	54.2	72.3	94.1	66.6	63.5	70.1	83.1	79.8
PolarNet [14]	L	71.0	74.7	28.2	85.3	90.9	35.1	77.5	71.3	58.8	57.4	76.1	96.5	71.1	74.7	74.0	87.3	85.7
Salsanext [15]	L	72.2	74.8	34.1	85.9	88.4	42.2	72.4	72.2	63.1	61.3	76.5	96.0	70.8	71.2	71.5	86.7	84.4
AMVNet [16]	L	76.1	**79.8**	32.4	82.2	86.4	**62.5**	81.9	75.3	**72.3**	**83.5**	65.1	**97.4**	67.0	**78.8**	74.6	90.8	87.9
Cylinder3D [2]	L	76.1	76.4	40.3	91.2	**93.8**	51.3	78.0	78.9	64.9	62.1	84.4	96.8	71.6	76.4	75.4	90.5	87.4
RPVNet [17]	L	77.6	78.2	43.4	92.7	93.2	49.0	**85.7**	80.5	66.0	66.9	84.0	96.9	73.5	75.9	76.0	90.6	88.9
PMF [18]	L+C	76.9	74.1	46.6	89.8	92.1	57.0	77.7	80.9	70.9	64.6	82.9	95.5	73.3	73.6	74.8	89.4	87.7
2D3DNet [19]	L+C	79.0	78.3	**55.1**	95.4	87.7	59.4	79.3	80.7	70.2	68.2	86.6	96.1	**74.9**	75.7	75.1	**91.4**	**89.9**
Baseline	L	76.2	75.3	43.5	95.3	91.2	54.5	78.9	72.8	62.1	70.0	83.2	96.3	73.2	74.2	74.9	88.1	85.9
2DPASS(Ours)	L	**79.4**	78.8	49.6	**95.6**	93.6	60.0	84.1	**82.2**	67.5	72.6	**88.1**	96.8	72.8	76.2	**76.5**	89.4	87.2

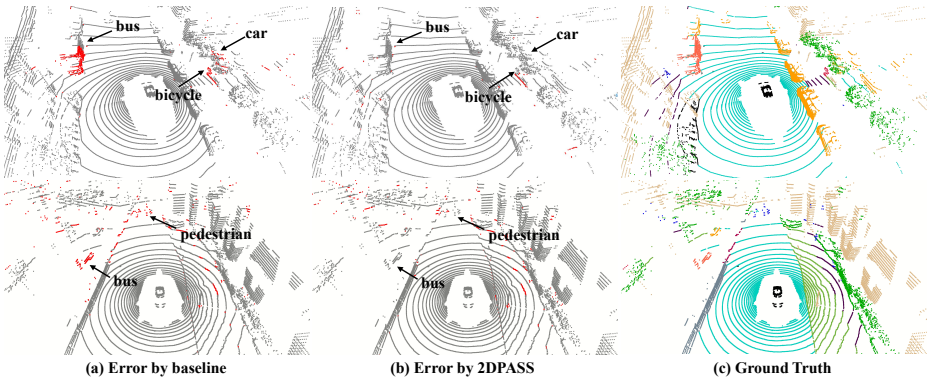


Fig. 2. Qualitative results of 2DPASS on the validation set of Nuscenes.

2DPASS (tiny) achieves the second-best performance in term of mIoU criterion while 60× speed faster than multi-sensor methods.

C.2 Concrete Results

In this section, we give our detailed results and visualization on the NuScenes dataset in Tab. 2 and Fig. 2 as a benchmark for future work. Also, we provide snapshots on three benchmarks in Fig. 3-5.

Rank	Participant team	mIOU (↑)	barrier (↑)	bicycle (↑)	bus (↑)	car (↑)	constr_vehicle (↑)	motorcycle (↑)	pedestrian (↑)	traffic_cone (↑)	trailer (↑)
1	SVQNet (SVQNet)	0.81	0.85	0.42	0.93	0.93	0.69	0.86	0.84	0.78	0.85
2	MIT HAN Lab (SPVCNN++)	0.81	0.86	0.43	0.92	0.92	0.76	0.76	0.83	0.77	0.87
3	2DPASS	0.81	0.82	0.55	0.92	0.92	0.73	0.86	0.79	0.72	0.85
4	DRINet++ (DRINet++: Efficient Voxel-as-p)	0.80	0.86	0.43	0.90	0.92	0.65	0.86	0.83	0.73	0.84
5	Uisee-FR (GU-Net)	0.80	0.85	0.33	0.87	0.91	0.74	0.85	0.81	0.78	0.88
6	2D3DNet (2D3DNet)	0.80	0.83	0.59	0.88	0.85	0.64	0.84	0.82	0.76	0.85
7	Kyber (AF2S3Net)	0.78	0.79	0.52	0.90	0.84	0.77	0.74	0.77	0.72	0.84
8	Cylinder3D++ (Cylinder3D++)	0.78	0.83	0.34	0.84	0.89	0.70	0.79	0.77	0.73	0.85
9	RH (CPFusion)	0.78	0.84	0.37	0.89	0.86	0.70	0.77	0.78	0.75	0.83
10	MIT HAN LAB (SPVNAS)	0.77	0.80	0.30	0.92	0.91	0.65	0.79	0.76	0.71	0.81

Fig. 3. Snapshot on NuScenes online server (2022-03-08).

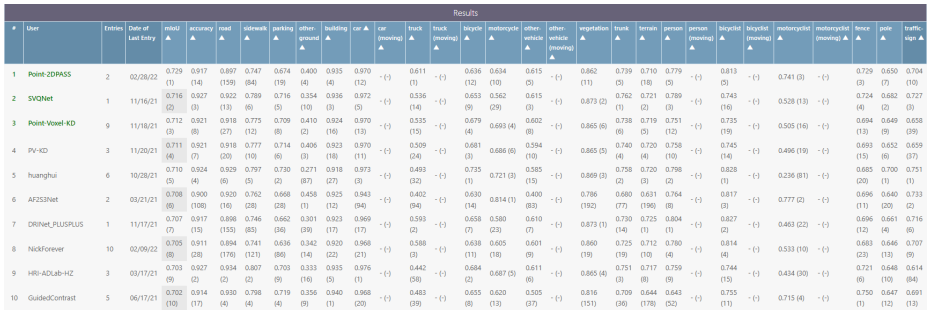


Fig. 4. Snapshot on SemanticKITTI online benchmark for single-scan competition (2022-03-08).

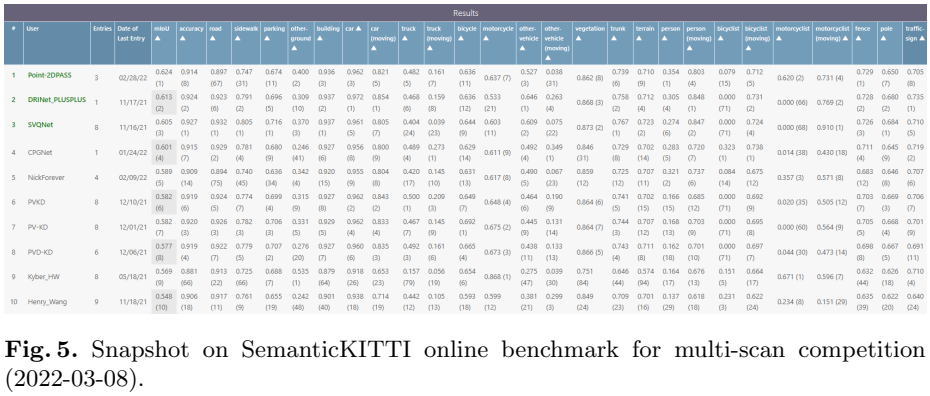


Fig. 5. Snapshot on SemanticKITTI online benchmark for multi-scan competition (2022-03-08).

References

1. Tang, H., Liu, Z., Zhao, S., Lin, Y., Lin, J., Wang, H., Han, S.: Searching efficient 3d architectures with sparse point-voxel convolution. In: European conference on computer vision, Springer (2020) 685–702
2. Zhou, H., Zhu, X., Song, X., Ma, Y., Wang, Z., Li, H., Lin, D.: Cylinder3d: An effective 3d framework for driving-scene lidar semantic segmentation. arXiv preprint arXiv:2008.01550 (2020)
3. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. NeurIPS Workshops (2014)
4. Huang, Z., Shen, X., Xing, J., Liu, T., Tian, X., Li, H., Deng, B., Huang, J., Hua, X.S.: Revisiting knowledge distillation: An inheritance and exploration framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 3579–3588
5. Jing Yang, Brais Martinez, A.B.G.T.: Knowledge distillation via softmax regression representation learning. In: ICLR2021. (2021)
6. Jaritz, M., Vu, T.H., Charette, R.d., Wirbel, E., Pérez, P.: xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 12605–12614
7. Vora, S., Lang, A.H., Helou, B., Beijbom, O.: Pointpainting: Sequential fusion for 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2020) 4604–4612
8. Yuan, Y., Huang, L., Guo, J., Zhang, C., Chen, X., Wang, J.: Ocnet: Object context network for scene parsing. arXiv preprint arXiv:1809.00916 (2018)
9. Milioto, A., Vizzo, I., Behley, J., Stachniss, C.: Rangenet++: Fast and accurate lidar semantic segmentation. In: Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS). (2019)
10. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 3431–3440
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
12. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
13. Cheng, R., Razani, R., Taghavi, E., Li, E., Liu, B.: Af2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2021) 12547–12556
14. Zhang, Y., Zhou, Z., David, P., Yue, X., Xi, Z., Gong, B., Foroosh, H.: Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 9601–9610
15. Cortinhal, T., Tzelepis, G., Aksoy, E.E.: Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds for autonomous driving. arXiv preprint arXiv:2003.03653 (2020)
16. Liong, V.E., Nguyen, T.N.T., Widjaja, S., Sharma, D., Chong, Z.J.: Amvnet: Assertion-based multi-view fusion network for lidar semantic segmentation. arXiv preprint arXiv:2012.04934 (2020)

17. Xu, J., Zhang, R., Dou, J., Zhu, Y., Sun, J., Pu, S.: Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 16024–16033
18. Zhuang, Z., Li, R., Jia, K., Wang, Q., Li, Y., Tan, M.: Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 16280–16290
19. Genova, K., Yin, X., Kundu, A., Pantofaru, C., Cole, F., Sud, A., Brewington, B., Shucker, B., Funkhouser, T.: Learning 3d semantic segmentation with only 2d image supervision. In: 2021 International Conference on 3D Vision (3DV), IEEE (2021) 361–372