ProtoTransfer: Cross-Modal Prototype Transfer for Point Cloud Segmentation —Supplementary Material—

Pin Tang[1] Hai-Ming Xu[2] Chao Ma[1]*

[1] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University

[2] Australian Institute for Machine Learning, University of Adelaide

{pin.tang,chaoma}@sjtu.edu.cn, hai-ming.xu@adelaide.edu.au

In this supplementary material, we provide more details to complement the manuscript, including implementation details in Sec. 1 and additional experimental results in Sec. 2.

1. Implementation details

This section presents more implementation details of the Proposed ProtoTransfer approach.

1.1. Model architecture

For fair comparison with the distillation-based method 2DPASS [10], we use the same model architecture as it. Two different encoder-decoder backbone are employed for 2D image and 3D point cloud.

For image, we apply ResNet-34 [4] as encoder to extract multi-scale image features. Then, we adpot a simple FCN [6] decoder which upsamples these multi-scale features to the resolution of 1/4 of the input size and concatenates them along the channel dimension. Finally, we can obtain the semantic segmentation of the input image by passing the concatenated feature map through a linear classifier.

For point cloud, we use the hierarchical encoder named SPVCNN [8] as in [10] to get the multi-scale point features. Note that each SPVCNN [8] encoder layer generates features for all LiDAR points. Hence, we simply concatenate these multi-scale point features in the decoder. Another linear classifier is utilized to get the semantic segmentation of the input LiDAR points.

1.2. Point-to-pixel mapping mechanism

We use the same point-to-pixel mapping mechanism as in PMF [42] and 2DPASS [35]. Formally, given a point cloud $P = \{p_i\}_{i=1}^{N}$, we can project each point $p_i := \{x_i, y_i, z_i\}$ to the corresponding pixel coordinate in the image plane as

$$[u_i, v_i, 1]^T = \frac{1}{z_i} \times K \times T \times [x_i, y_i, z_i, 1]^T,$$

where $K \in \mathbb{R}^{3\times 4}$ and $T \in \mathbb{R}^{4\times 4}$ are the camera intrinsic and extrinsic matrices respectively, which are sensor configurations provided by the dataset. Then, we obtain the point-to-pixel one-to-one correspondence mapping $M^{\mathrm{p2p}} = \{(\lfloor u_i \rfloor, \lfloor v_i \rfloor)\}_{i=1}^{N}$ where $\lfloor \cdot \rfloor$ is the round down operation. Since only part of the coordinates are located within the image plane, we construct a mask $M^{\mathrm{3d}} \in [0,1]^N$, where the above coordinates located within the image plane are assigned to 1 and the remaining ones are assigned to 0. Next, we obtain the coordinates of matched pixels by $M^{\mathrm{2d}} = M^{\mathrm{p2p}}[M^{\mathrm{3d}}]$ and further get the matched pixel features $F^{\mathrm{2d\_m}} = F^{\mathrm{2d}}[M^{\mathrm{2d}}]$ and point features $F^{\mathrm{3d\_m}} = F^{\mathrm{3d}}[M^{\mathrm{3d}}]$ through the tensor indexing operation. $F^{\mathrm{3d\_m}}$ and $F^{\mathrm{2d\_m}}$ extracted using M^{3d} and M^{2d} are naturally paired and ordered since M^{p2p} ensures the correspondence between points and pixels. Thus, $F^{\mathrm{3d\_m}}$ and $F^{\mathrm{2d\_m}}$ are in the same shape and can be concatenated as in Eq. (1).

1.3. Training and inference

During training, point clouds and images are differently pre-processed. Specifically, for the image modal, images are firstly randomly cropped and resized to 480 × 320 as done in 2DPASS [10]. Then, color jittering and horizontal flipping are also used to augment the input images. For the point cloud modal, the whole cloud frame is utilized, and several augmentations operations are included, i.e., random point dropout, global rotation and translation, global scaling, and flipping. For the optimizer, the learning rate is set as 0.24 and a cosine function learning strategy is utilized to gradually decay the learning rate. The batch size is set to 8. The model is trained for 64 epochs on SemanticKITTI and 80 epochs on nuScenes.

During inference, the network only takes the point clouds as input while the image segmentation branch and

[*] Corresponding author.

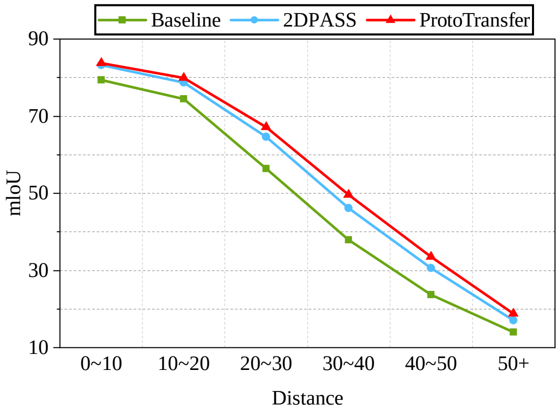


Figure A. Distance-based evaluation on nuScenes Lidarseg validation set.

Method	mIoU
baseline	76.2
2DPASS [10]	79.4
ProtoTransfer (Proto. head)	79.6
ProtoTransfer (learned head)	80.5

Table A. Abaltion study of segmentation head on nuScenes Lidarseg validation set.

the multi-modal fusion branch are dropped. Moreover, test-time augmentation is used for performance boosting. Both training and inference are conducted with 8 NVIDIA Tesla V100 GPUs.

2. Additional Experiments

2.1. Distance-based evaluation

In this part, we study how segmentation is affected by distance of the points to the ego-vehicle on nuScenes validation set. As illustrated in Fig. A, the results of all methods get worse by increasing the distance as point clouds become sparser. However, as can be observed, our ProtoTransfer improves the performance in 20-50m significantly, effectively alleviating this trend. We attribute this performance gain to the fully exploited and transferred multi-modal knowledge.

2.2. Generality

To evaluate the generality of our ProtoTransfer, we apply it to MinkowskiNet [8]. MinkowskiNet achieves a mIoU gain of +2.6% (from 76.0% to 78.6%) on the nuScenes validation set, affirming the generality of our ProtoTransfer.

2.3. Use prototype bank as a segmentation head

Although the prototype bank in our method is used as a bridge for multi-modal knowledge transfer, it can also be served as a non-learnable segmentation head as proposed in [12]. Therefore, we further use the class-wise prototype bank in our model as a segmentation head by assigning each LiDAR point to its closest prototype. The segmentation results are given in Tab. A. The ProtoTransfer using prototype bank as segmentation head achieves 79.6% mIoU, which surpasses our baseline and the former stat-of-the-art 2DPASS [10], showing a solid discriminative ability of the prototype bank.

2.4. Results on nuScenes

Tab. B presents the results of our ProtoTransfer and previously published methods on the nuScenes Lidarseg validation set. We can find that our ProtoTransfer outperforms 2DPASS, the former state-of-the-art, by 1.1% in terms of mIoU. Besides, ProtoTransfer achieves the best or comparable performance on classes of small sizes, such as bicycle, motorcycle and pedestrian, showing the superiority of our methods in multi-modal fusion and knowledge transfer.

The screenshot of the test results on the online leaderboard of nuScenes Lidarseg Challenge is shown in Fig. B, where our ProtoTransfer ranks 2nd.

References

[1] Ran Cheng, Ryan Razani, Ehsan Taghavi, Enxu Li, and Bingbing Liu. 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*, pages 12547–12556, 2021.

[2] Tiago Cortinhal, George Tzelepis, and Eren Erdal Aksoy. Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds for autonomous driving. *arXiv preprint arXiv:2003.03653*, 2020.

[3] Kyle Genova, Xiaoqi Yin, Abhijit Kundu, Caroline Pantofaru, Forrester Cole, Avneesh Sud, Brian Brewington, Brian Shucker, and Thomas Funkhouser. Learning 3d semantic segmentation with only 2d image supervision. In *2021 International Conference on 3D Vision (3DV)*, pages 361–372. IEEE, 2021.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

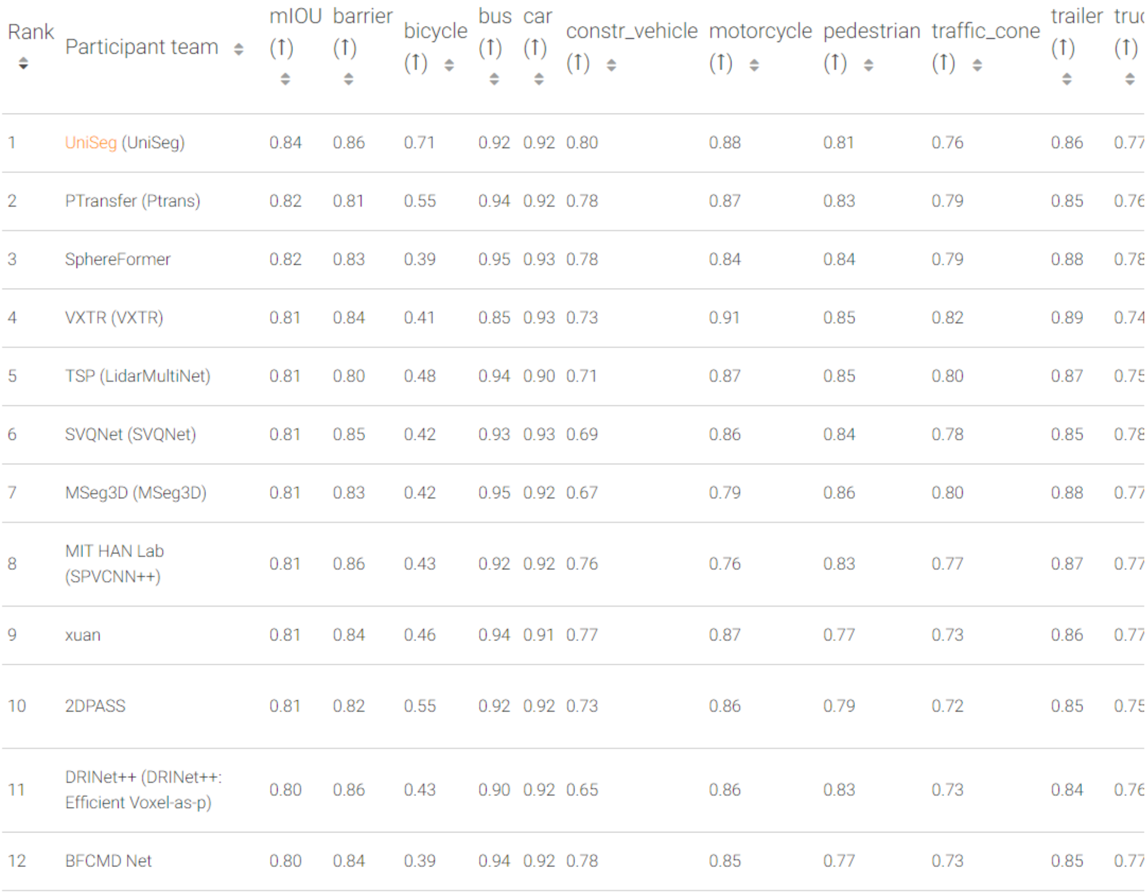
[5] Venice Erin Liong, Thi Ngoc Tho Nguyen, Sergi Widjaja, Dhananjai Sharma, and Zhuang Jie Chong. Amvnet: Assertion-based multi-view fusion network for lidar semantic segmentation. *arXiv preprint arXiv:2012.04934*, 2020.

[6] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[7] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Rangenet++: Fast and accurate lidar semantic segmentation. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4213–4220. IEEE, 2019.

Methods	Input	**mIoU**	barrier	bicycle	bus	car	construction	motorcycle	pedestrian	traffic-cone	trailer	truck	driveable	other	sidewalk	terrain	manmade	vegetation
$(AF)^2$-S3Net [1]	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++ [7]	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 [11]	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 [2]	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 [5]	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 [13]	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 [9]	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 [14]	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 [3]	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**
2DPASS [10]	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
Baseline [10]	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
ProtoTransfer [ours]	L	**80.5**	78.4	54.0	95.5	93.0	60.7	**89.0**	**83.4**	69.8	76.7	**88.3**	96.8	74.6	76.0	75.6	89.4	87.0

Table B. Results of our proposed method and published state-of-the-art LiDAR Segmentation methods on nuScenes Lidarseg validation set, where L and C respectively denote LiDAR and camera. The bold numbers indicate the best results.



Rank	Participant team	mIOU (↑)	barrier (↑)	bicycle (↑)	bus (↑)	car (↑)	constr_vehicle (↑)	motorcycle (↑)	pedestrian (↑)	traffic_cone (↑)	trailer (↑)	truc (↑)
1	UniSeg (UniSeg)	0.84	0.86	0.71	0.92	0.92	0.80	0.88	0.81	0.76	0.86	0.77
2	PTransfer (Ptrans)	0.82	0.81	0.55	0.94	0.92	0.78	0.87	0.83	0.79	0.85	0.76
3	SphereFormer	0.82	0.83	0.39	0.95	0.93	0.78	0.84	0.84	0.79	0.88	0.78
4	VXTR (VXTR)	0.81	0.84	0.41	0.85	0.93	0.73	0.91	0.85	0.82	0.89	0.74
5	TSP (LidarMultiNet)	0.81	0.80	0.48	0.94	0.90	0.71	0.87	0.85	0.80	0.87	0.75
6	SVQNet (SVQNet)	0.81	0.85	0.42	0.93	0.93	0.69	0.86	0.84	0.78	0.85	0.78
7	MSeg3D (MSeg3D)	0.81	0.83	0.42	0.95	0.92	0.67	0.79	0.86	0.80	0.88	0.77
8	MIT HAN Lab (SPVCNN++)	0.81	0.86	0.43	0.92	0.92	0.76	0.76	0.83	0.77	0.87	0.77
9	xuan	0.81	0.84	0.46	0.94	0.91	0.77	0.87	0.77	0.73	0.86	0.77
10	2DPASS	0.81	0.82	0.55	0.92	0.92	0.73	0.86	0.79	0.72	0.85	0.75
11	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	0.76
12	BFCMD Net	0.80	0.84	0.39	0.94	0.92	0.78	0.85	0.77	0.73	0.85	0.77

Figure B. Screen shot of nuScenes Lidarseg Challenge leaderboard on online server (2023-02-14). The 2nd palce "PTransfer (Ptrans)" is ours.

[8] 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.

[9] Jianyun Xu, Ruixiang Zhang, Jian Dou, Yushi Zhu, Jie Sun, and Shiliang Pu. 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*, pages 16024–16033, 2021.

[10] Xu Yan, Jiantao Gao, Chaoda Zheng, Chao Zheng, Ruimao Zhang, Shuguang Cui, and Zhen Li. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In

European Conference on Computer Vision, pages 677–695. Springer, 2022.

[11] Yang Zhang, Zixiang Zhou, Philip David, Xiangyu Yue, Zerong Xi, Boqing Gong, and Hassan Foroosh. 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*, pages 9601–9610, 2020.

[12] Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2582–2593, 2022.

[13] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9939–9948, 2021.

[14] Zhuangwei Zhuang, Rong Li, Kui Jia, Qicheng Wang, Yuanqing Li, and Mingkui Tan. Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16280–16290, 2021.