3D Object Detection with a Self-supervised Lidar Scene Flow Backbone Supplementary Material

Emeç Erçelik[1⋆], Ekim Yurtsever[2*], Mingyu Liu[1,3], Zhijie Yang[1], Hanzhen Zhang[1], Pınar Topçam[1], Maximilian Listl[1], Yılmaz Kaan Çaylı[1], and Alois Knoll[1]

[1] Chair of Robotics, Artificial Intelligence and Real-time Systems, Technical University of Munich, 85748 Garching b. München, Germany
{emec.ercelik, mingyu.liu, zhijie.yang, hanzhen.zhang, pinar.topcam, maximilian.listl, kaan.cayl }@tum.de, knoll@in.tum.de
[2] Ohio State University, Columbus, OH 43212, USA
yurtsever.2@osu.edu
[3] Tongji University, 201804, Shanghai, China

1 Point-GNN [5] KITTI Validation Results with R_{11} Metric

Point-GNN paper [4] provides its KITTI validation set scores with $AP_{R_{11}}$ metric. Therefore, we show the comparison between the self-supervised Point-GNN, the baseline Point-GNN, and the previous state-of-the-art using the $AP_{R_{11}}$ in Table 1. The baseline Point-GNN $AP_{R_{11}}$ scores are taken from the Point-GNN paper [5]. Our self-supervised Point-GNN outperforms the baseline on all difficulty levels.

Car (IoU=0.7)	3D AP		
Method	Easy	Mod	Hard
VoxelNet[8]	81.97	65.46	62.85
F-PointNet[3]	83.76	70.92	63.65
AVOD[1]	84.41	74.44	68.65
SECOND[6]	87.43	76.48	69.10
PointPillars[2]	-	77.98	-
Point-GNN [5]	87.89	78.34	77.38
Self-supervised Point-GNN	**88.32**	**78.66**	**77.92**

Table 1. Self-supervised Point-GNN 3D $AP_{R_{11}}$ car results on KITTI validation set.

⋆ Authors contributed equally.

2 Point-GNN [5] Alternating Training Ablation

We compare Point-GNN 3D detector pre-trained with and without alternating training in Table 2 to justify our training strategy. The alternating training strategy repeats the self-supervised backbone training followed by the supervised 3D detection fine-tuning pair several times. We repeat the pair-wise training twice to get alternating training results. The *Self-sup Point-GNN* in Table 2 applies the pair-wise training (self-supervised followed by supervised) only once. For comparison purposes, we also include Point-GNN baseline results in this table. The alternating training strategy enhances the car class 3D detection precision for more than 2% AP in the moderate difficulty level.

Car (IoU=0.7)	3D AP			BEV AP		
Method	Easy	Mod	Hard	Easy	Mod	Hard
Point-GNN baseline [5]	90.44	82.12	77.70	93.03	89.31	86.86
Self-sup. Point-GNN	91.05	82.35	77.80	93.43	89.59	87.03
Self-sup. Point-GNN w/ alternating tr.	**91.43**	**82.85**	**80.12**	**93.55**	**89.79**	**87.23**

Table 2. Ablation results for applying the self-supervised training alternating with the 3D detection fine-tuning.

3 Ablation on the nuScenes Dataset: Limited Labeled Data and Alternating Training

We further investigate performance of our self-supervised scene flow pre-training method with CenterPoint[7], PointPillars[2], and SSN[9] 3D detectors on nuScenes dataset.

Fig. 1 shows the CenterPoint mAP and NDS 3D detection scores for training from scratch and using our pre-training method. We show the percentage of the training set used for supervised training in the x-axes. The blue lines indicate the baseline CenterPoint trained from scratch. We show the self-supervised CenterPoint results with orange lines (Ours_a) without alternating training and green lines (Ours_b) with alternating training. As seen, our pre-training enhances the 3D detection scores compared to the baseline, and the alternating training strategy further improves the results.

We show the 3D detection results (mAP and NDS) of the baseline and our self-supervised PointPillars in Fig 2. Blue lines give the baseline PointPillars trained with the given percentages of the labeled data. Ours_a and Ours_b belong to our self-supervised PointPillars without and with alternating training, respectively. The PointPillars obtained apparent improvements over the baseline based on our self-supervised pre-training approach.

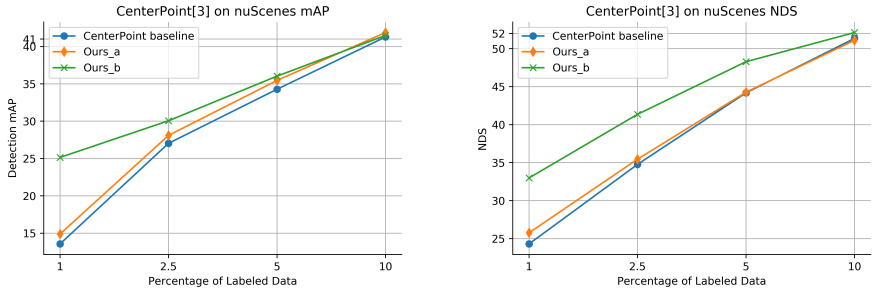

Fig. 1. CenterPoint [7] 3D detection ablation results on the nuScenes dataset. All the CenterPoint versions are trained with the given percentage of the labeled data in the x-axes. The baseline (blue line) indicates training from scratch. Ours_a (orange) and Ours_b (green) are without and with alternating training. Our self-supervised pre-training enhances nuScenes 3D detection scores (mAP and NDS) compared to the baseline.

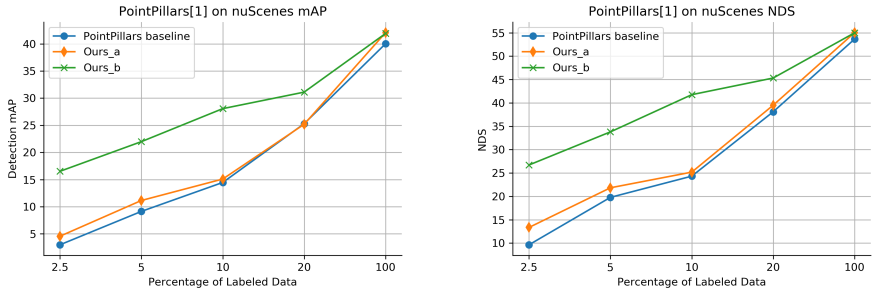

Fig. 2. PointPillars[2] 3D detection ablation results on the nuScenes dataset. All the PointPillars versions are trained with a part of labeled data shown in the x-axis. The baseline (blue) was trained with labeled data from scratch. Our self-supervised approaches Ours_a without alternating training and Ours_b with alternating training outperform the baseline on mAP and NDS metrics significantly.

We draw similar curves for the SSN 3D detector [4] trained with a part of annotated nuScenes 3D detection training split in Fig. 3. Blue lines are for the baseline SSN. Ours_a and Ours_b show the SSN 3D detection scores pre-trained without and with alternating training strategy. Our alternating self-supervised pre-training strategy improves the 3D mAP and NDS with a large margin compared to the baseline.

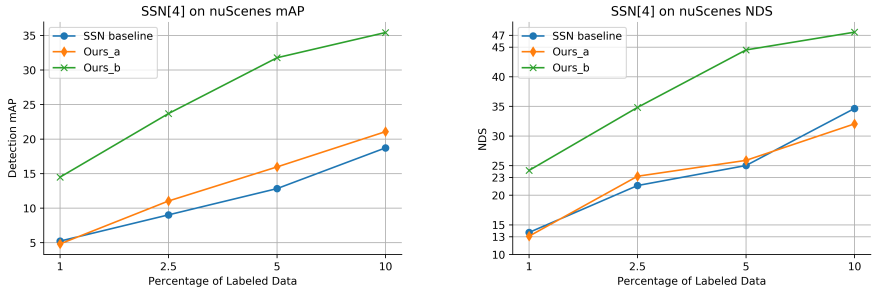

Fig. 3. SSN[9] 3D detection ablation results on the nuScenes dataset. We trained the baseline (blue) with the labeled data from scratch. Ours_a (without alternating training) and Ours_b (with alternating training) are pre-trained with our self-supervised scene flow method and then trained with the labeled data. All the SSN versions are trained with the indicated percentage of the labeled data in the x-axis. Our pre-trained SSN outperforms the baseline on mAP and NDS metrics.

4 PointPillars [2] Ablation on KITTI: Limited Labeled Data

We compare our pre-training with the baseline using PointPillars trained with a small part of training split in Fig. 4 in addition to the PointPillars 3D detection results on nuScenes dataset given in the main paper. Blue and orange lines are for the baseline and our pre-trained PointPillars, respectively. The y-axis shows the mAP scores for all three classes on moderate difficulty level. As seen in the figure, our self-supervised pre-training provides better 3D detection mAP than training from scratch.

5 Conclusion

Further ablation study results in this supplementary material suggest that our self-supervised scene flow pre-training helps achieve better 3D detection accuracy for several 3D detectors on KITTI and nuScenes datasets. Learning motion representations with the scene flow pre-training is especially useful when

[4] https://github.com/open-mmlab/mmdetection3d/

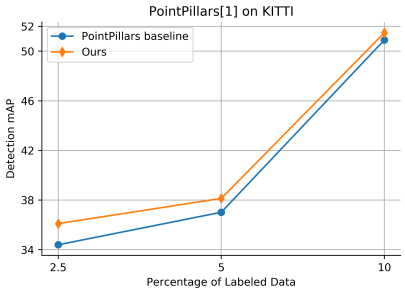

Fig. 4. 3D detection results using PointPillars [2] trained with a small part of KITTI data. Ours (orange) is pre-trained with our self-supervised scene flow. Baseline (blue) is trained with the labeled data from scratch. Our pre-trained PointPillars outperforms the baseline trained with the limited labeled data.

the annotated dataset is limited. Moreover, our alternating training strategy contributes to learning the relation between motion and geometric point cloud features, resulting in better 3D detection.

References

1. Ku, J., Mozifian, M., Lee, J., Harakeh, A., Waslander, S.L.: Joint 3d proposal generation and object detection from view aggregation. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1–8. IEEE (2018)
2. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12697–12705 (2019)
3. Qi, C.R., Liu, W., Wu, C., Su, H., Guibas, L.J.: Frustum pointnets for 3d object detection from rgb-d data. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 918–927 (2018)
4. Shi, S., Wang, Z., Shi, J., Wang, X., Li, H.: From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. IEEE transactions on pattern analysis and machine intelligence **43**(8), 2647–2664 (2020)
5. Shi, W., Rajkumar, R.: Point-gnn: Graph neural network for 3d object detection in a point cloud. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1711–1719 (2020)
6. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors **18**(10), 3337 (2018)
7. Yin, T., Zhou, X., Krahenbuhl, P.: Center-based 3d object detection and tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11784–11793 (2021)
8. Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4490–4499 (2018)
9. Zhu, X., Ma, Y., Wang, T., Xu, Y., Shi, J., Lin, D.: Ssn: Shape signature networks for multi-class object detection from point clouds. In: European Conference on Computer Vision. pp. 581–597. Springer (2020)