# Supplementary Material: Equivariant Spatio-Temporal Self-Supervision for LiDAR Object Detection

Deepti Hegde[1], Suhas Lohit[2], Kuan-Chuan Peng[2], Michael J. Jones[2], and Vishal M. Patel[1]

[1] Johns Hopkins University, Baltimore, MD 21218, USA
[2] Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139, USA
{dhegde1,vpatel36}@jhu.edu, {slohit,kpeng,mjones}@merl.com

## A Overview

We present supplementary material for this paper. In Section A.1, we include additional quantitative results of training **E-SSL$^{3D}$** on the KITTI-360 [4] and Waymo Open Datasets [5] for fine-tuning on the SECOND dataset. In Section A.2.

### A.1 Training SECOND with E-SSL$^{3D}$

We present additional experimental results of training **E-SSL$^{3D}$** for the purpose of fine-tuning the object detector SECOND [7] on the KITTI object detection dataset [2]. We perform pre-training with two datasets, the Waymo Open Dataset (WOD) [5] and the KITTI-360 dataset [4]. The WOD pre-training dataset consists of $100k$ samples cropped to a front field-of-view to be consistent with the main experimental settings. In order to mitigate the distribution gap between datasets, we train the networks on $\{x, y, z\}$ coordinates, leaving out intensity. We demonstrate detection performance on 3 different splits of data, $\{5\%, 20\%, 100\%\}$. We compare the performance of our approach with that of PointContrast [6] and STRL [3]. As the weights of ALSO [1] trained on WOD are unavailable, we forgo this comparison. We follow the same experimental settings mentioned in the main text.

In Table 2, we present the results of SECOND pre-trained with the WOD and fine-tuned on three data splits from KITTI. We consistently perform best or second best across all splits. Particularly in the 5% and 100% data splits, we out-perform comparative methods by a large margin. In Table 1 we present the results of SECOND pre-trained with the KITTI-360 dataset and fine-tuned on three data splits from KITTI. We perform best or second best across most categories, and outperform the invariant counterpart STRL [3].

### A.2 Fine-tuning VoxelRCNN on Waymo

In Table 3 we include results on pre-training and fine-tuning VoxelRCNN on the Waymo Open Dataset dataset. We demonstrate the performance of VoxelRCNN

**Table 1:** 3D object detection with SECOND [8] pre-trained on KITTI-360 [4] and fine-tuned on KITTI [2] under different data splits. Each result is an average over 3 fixed subsets of the dataset. We report 3D average precision for 3 categories as well as the mean average precision over 40 recall positions. The best and second best performance is marked in **bold** and underline, respectively.

| Split | Method | average precision (AP) (%) | | | | | | | | | mAP (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Car | | | Pedestrian | | | Cyclist | | | |
| | | easy | moderate | hard | easy | moderate | hard | easy | moderate | hard | |
| 5% | No pre-training | 86.40 | 73.01 | 67.79 | 42.76 | **38.26** | **34.42** | 64.13 | 45.44 | 42.70 | 54.99 |
| | PointContrast | 86.93 | 73.48 | 69.30 | 41.11 | 37.67 | 34.34 | 64.31 | 47.26 | 44.21 | 55.40 |
| | STRL | 86.61 | 72.90 | 68.37 | 39.24 | 35.46 | 31.80 | 59.90 | 42.52 | 39.66 | 52.94 |
| | ALSO | 86.80 | **75.56** | **72.88** | 41.92 | 36.90 | 34.02 | **74.72** | **58.99** | **55.42** | **59.69** |
| | **E-SSL$^{3D}$** | **87.53** | 74.96 | 70.47 | **43.30** | 37.71 | 34.40 | 74.02 | 54.13 | 50.91 | 58.60 |
| 20% | No pre-training | 87.64 | 77.12 | 73.13 | 50.36 | 45.70 | 41.29 | 75.26 | 54.99 | 51.24 | 61.86 |
| | PointContrast | 88.03 | 77.49 | 73.15 | 50.45 | 45.97 | 41.21 | 74.07 | 54.48 | 50.65 | 61.72 |
| | STRL | 87.97 | 77.50 | 73.25 | 51.68 | 46.82 | 42.00 | 72.49 | 53.89 | 50.08 | 61.74 |
| | ALSO | 88.73 | **79.44** | **76.32** | **52.46** | **47.71** | **44.28** | **81.74** | **65.50** | **61.31** | **66.39** |
| | **E-SSL$^{3D}$** | **89.83** | 78.70 | 75.93 | 51.65 | 46.10 | 42.40 | 80.83 | 62.92 | 59.13 | 65.28 |
| 100% | No pre-training | 90.05 | 81.03 | 78.09 | 53.97 | 49.19 | 44.06 | 80.59 | 63.54 | 59.64 | 66.69 |
| | PointContrast | 88.40 | 80.50 | 76.44 | 53.02 | 48.34 | 44.01 | 80.05 | 62.32 | 58.61 | 65.74 |
| | STRL | 90.37 | 81.11 | 78.21 | **59.55** | **53.61** | **48.08** | 78.72 | 62.47 | 57.92 | 67.78 |
| | ALSO | **90.44** | **81.66** | **78.83** | 56.41 | 51.91 | 47.53 | 84.49 | 67.65 | 63.53 | 69.18 |
| | **E-SSL$^{3D}$** | 90.14 | 81.64 | 78.61 | 57.53 | 53.05 | 47.72 | **85.40** | **69.54** | **64.69** | **69.81** |

pre-trained on the Waymo Open Dataset and fine-tuned on 5% of annotated samples from Waymo. We show that the best performance is achieved when the detector is pre-trained using **E-SSL$^{3D}$**, a trend that is observed across categories.

**Table 2:** 3D object detection with SECOND [8] pre-trained on the Waymo Open Dataset [5] and fine-tuned on KITTI [2] under different data splits. Each result is an average over 3 fixed subsets of the dataset. We report 3D average precision for 3 categories as well as the mean average precision over 40 recall positions. The best and second best performance is marked in **bold** and underline, respectively.

| Split | Method | average precision (AP) (%) | | | | | | | | | mAP (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Car | | | Pedestrian | | | Cyclist | | | |
| | | easy | moderate | hard | easy | moderate | hard | easy | moderate | hard | |
| 5% | No pre-training | **86.40** | 73.01 | 67.79 | 42.76 | 38.26 | 34.42 | 64.13 | 45.44 | 42.70 | 54.99 |
| | PointContrast | 86.32 | 73.14 | 69.73 | 42.78 | 36.93 | 33.47 | 70.51 | 49.76 | 46.77 | 56.60 |
| | STRL | 86.23 | 72.19 | 68.67 | 43.13 | 38.07 | 34.62 | 70.43 | 50.80 | 47.63 | 56.86 |
| | **E-SSL$^{3D}$** | 86.30 | **73.34** | **69.91** | **45.36** | **39.57** | **35.86** | **72.44** | **51.51** | **48.39** | **58.07** |
| 20% | No pre-training | 87.64 | 77.12 | 73.13 | 50.36 | 45.70 | 41.29 | 75.26 | 54.99 | 51.24 | 61.86 |
| | PointContrast | **90.42** | **79.09** | **76.13** | **52.18** | **46.68** | **42.91** | **80.87** | **63.75** | **59.77** | **65.75** |
| | STRL | 89.64 | 78.44 | 75.53 | 50.64 | 45.31 | 41.58 | 79.22 | 61.59 | 57.92 | 64.43 |
| | **E-SSL$^{3D}$** | 89.71 | 78.44 | 75.67 | 50.83 | 45.77 | 41.94 | 80.29 | 62.64 | 58.91 | 64.91 |
| 100% | No pre-training | 90.05 | 81.03 | 78.09 | 53.97 | 49.19 | 44.06 | 80.59 | 63.54 | 59.64 | 66.69 |
| | PointContrast | 88.06 | 78.98 | 76.05 | 52.85 | 47.18 | 42.29 | 79.04 | 60.80 | 56.85 | 64.68 |
| | STRL | 90.04 | 80.83 | 77.66 | 53.44 | 48.39 | 43.13 | 78.87 | 59.88 | 55.87 | 65.35 |
| | **E-SSL$^{3D}$** | **90.38** | **81.31** | **78.19** | **55.76** | **50.61** | **46.08** | **81.49** | **63.59** | **59.69** | **67.46** |

**Table 3:** Detection performance of VoxelRCNN for the “Cyclist" category pre-trained on the Waymo dataset and fine-tuned on 5% of Waymo data. Precision is reported using official Waymo evaluation metrics.

| Method | Level 1 | | | Level 2 | | |
|---|---|---|---|---|---|---|
| | AP | APH | APL | AP | APH | APL |
| No pre-training | 21.67 | 21.14 | 21.67 | 20.59 | 20.09 | 20.59 |
| PointContrast | 21.39 | 20.82 | 21.39 | 20.33 | 19.78 | 20.33 |
| STRL | 19.45 | 18.68 | 19.45 | 18.48 | 17.75 | 18.48 |
| **E-SSL$^{3D}$** | **22.99** | **22.33** | **22.99** | **21.86** | **21.23** | **21.86** |

## References

1. Boulch, A., Sautier, C., Michele, B., Puy, G., Marlet, R.: ALSO: Automotive lidar self-supervision by occupancy estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13455–13465 (2023)
2. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
3. Huang, S., Xie, Y., Zhu, S.C., Zhu, Y.: Spatio-temporal self-supervised representation learning for 3d point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6535–6545 (2021)
4. Liao, Y., Xie, J., Geiger, A.: KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. Pattern Analysis and Machine Intelligence (PAMI) (2022)
5. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2446–2454 (2020)
6. Xie, S., Gu, J., Guo, D., Qi, C.R., Guibas, L., Litany, O.: PointContrast: Unsupervised pre-training for 3D point cloud understanding. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. pp. 574–591. Springer (2020)
7. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors **18**(10), 3337 (2018)
8. Yan, Y., Mao, Y., Li, B.: SECOND: Sparsely embedded convolutional detection. Sensors **18**(10) (2018). https://doi.org/10.3390/s18103337, https://www.mdpi.com/1424-8220/18/10/3337