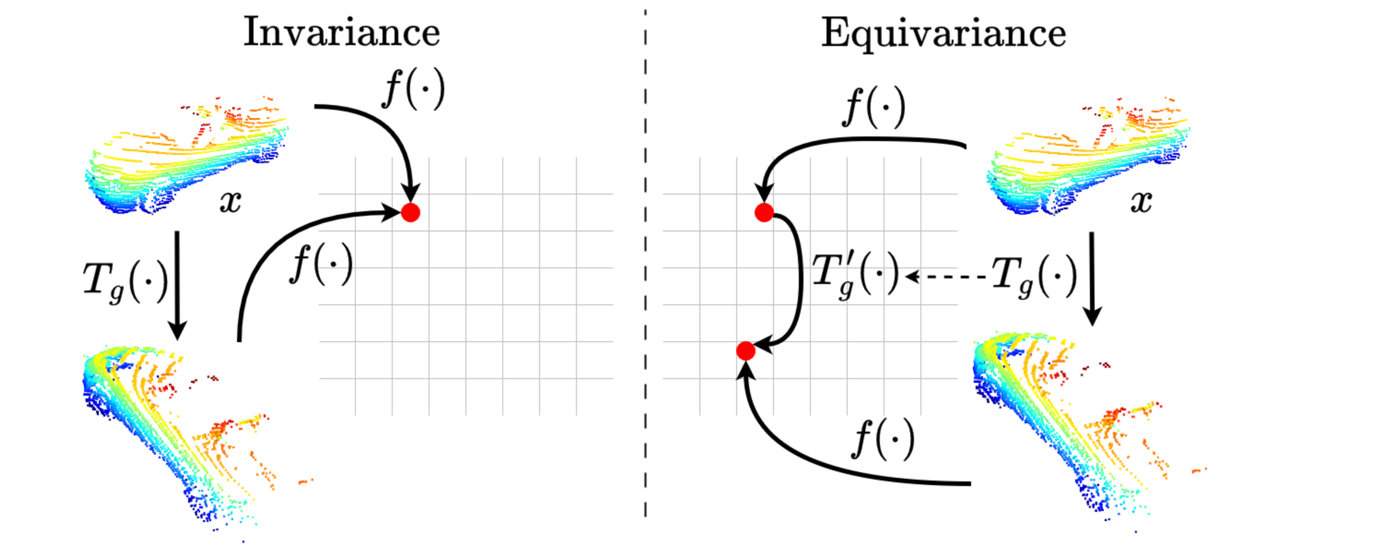
Equivariant Spatio-Temporal Self-Supervision for LiDAR Object Detection

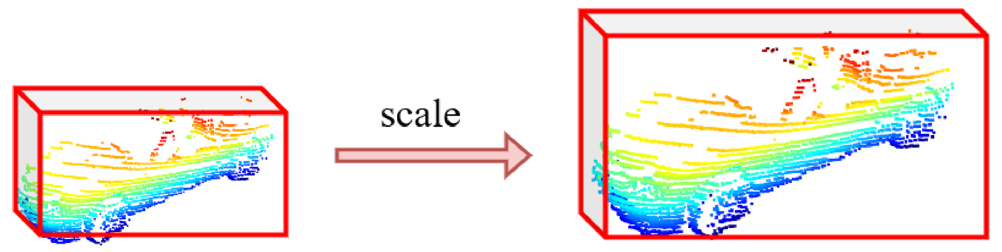
Deepti Hegde[1]*, Suhas Lohit[2], Kuan-Chuan Peng[2], Michael J. Jones[2], Vishal M. Patel[1]
[1]Johns Hopkins University, [2]Mitsubishi Electric Research Laboratories
* Work done while an intern at MERL

Self-supervised learning on point clouds

Discriminative features can be learned without any labeled data by encouraging invariance and equivariance to input transformations



Equivariant features are designed to retain information of the transformation needed for localization



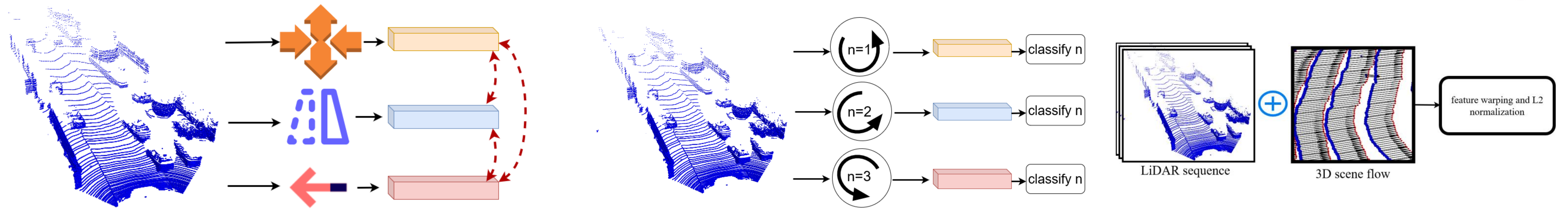
Bounding box scales with object

- PointContrast[1] encourages equivariance to spatial transformations through a contrastive learning objective and does not consider temporal transformations
- By using sequences of LiDAR frames and estimated 3D scene flow[4] we consider naturally occurring temporal transformations in addition to spatial ones
- STRL[3] encourages spatio-temporal **invariance** to learn effective representations

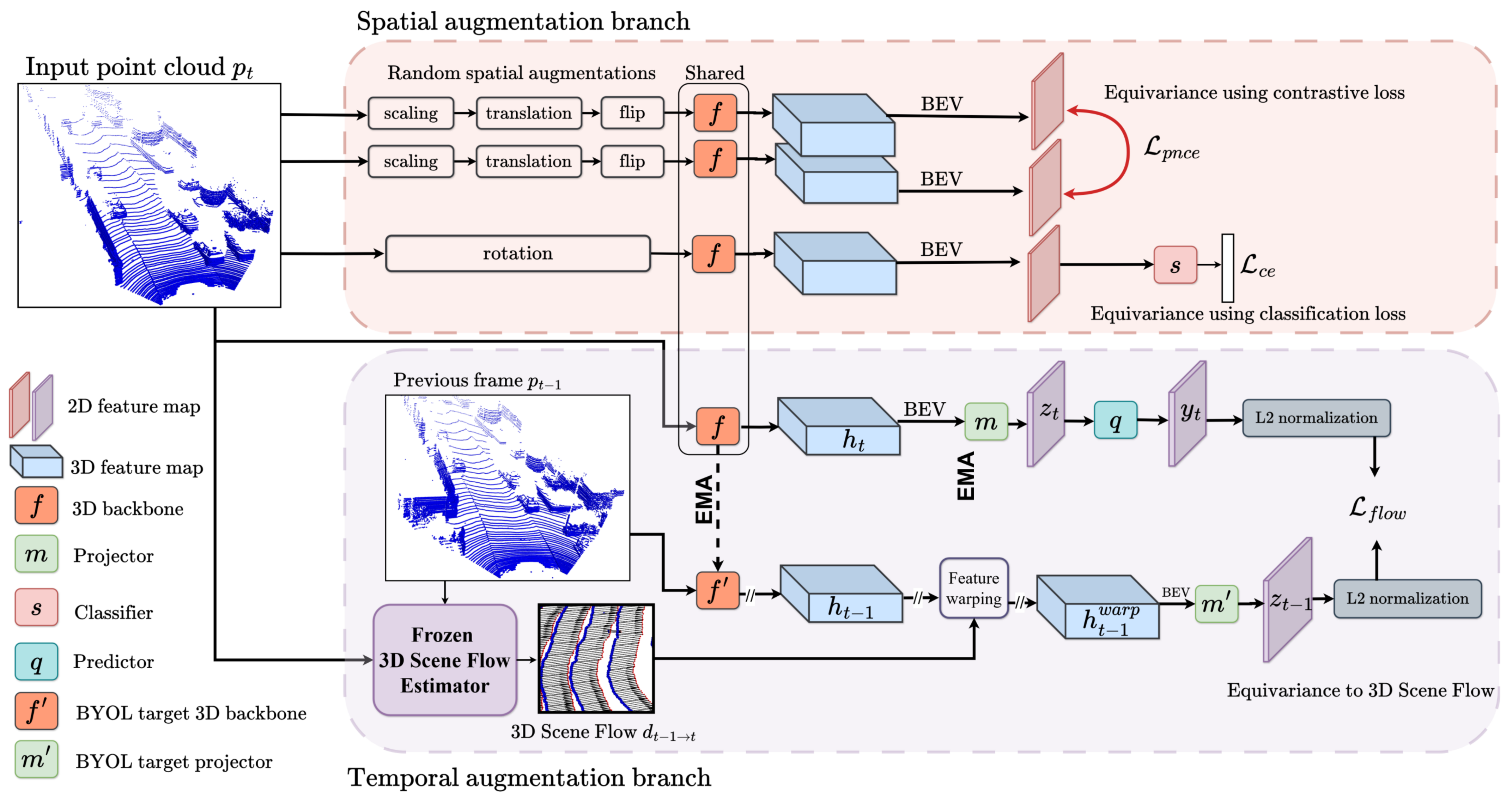
Equivariant self-supervised learning in space and time

We train the network to be **equivariant** to both spatial and temporal transformations through three loss objectives.

1. Equivariance to scaling, translation, and random scene flip through contrastive training
2. Equivariance to n-fold rotation through classification
3. Enforcing temporal equivariance with 3D scene flow



Pretraining point cloud feature extractors using E-SSL^{3D}



Results

We use the KITTI-360 and Waymo datasets for pre-training and demonstrate good performance on the downstream task of **3D object detection** with VoxelRCNN.

Split	Method	Car easy	Car moderate	Car hard	Pedestrian easy	Pedestrian moderate	Pedestrian hard	Cyclist easy	Cyclist moderate	Cyclist hard	mAP (%)
5%	No pre-training	88.89	<u>79.21</u>	75.55	**57.50**	**49.84**	<u>44.27</u>	78.92	59.73	55.97	65.54
	PointContrast	**89.94**	<u>79.21</u>	**76.12**	56.13	48.13	43.01	77.98	58.92	55.20	64.96
	STRL	89.30	78.92	<u>75.94</u>	55.68	48.13	42.73	73.98	56.85	53.26	63.87
	ALSO	<u>89.74</u>	**79.37**	75.91	<u>56.33</u>	<u>49.79</u>	**44.77**	<u>82.84</u>	<u>64.09</u>	<u>60.16</u>	**67.00**
	E-SSL^{3D}	88.79	78.93	75.41	56.02	48.55	43.19	**82.85**	**64.40**	**60.53**	<u>66.52</u>
20%	No pre-training	91.99	82.10	79.40	56.09	49.29	44.26	85.24	67.55	63.13	68.78
	PointContrast	92.23	82.25	79.57	57.33	50.74	45.43	84.16	66.74	62.28	68.97
	STRL	91.97	82.07	79.41	57.40	50.85	45.38	<u>86.36</u>	68.64	64.23	69.59
	ALSO	<u>92.46</u>	**82.44**	<u>79.77</u>	<u>60.57</u>	<u>53.21</u>	<u>48.61</u>	86.22	<u>69.88</u>	<u>65.40</u>	<u>70.95</u>
	E-SSL^{3D}	**92.67**	<u>82.42</u>	**79.89**	**60.72**	**53.94**	**49.19**	**88.04**	**71.40**	**66.36**	**71.63**
100%	No pre-training	<u>92.45</u>	**83.00**	<u>80.20</u>	**62.41**	**55.89**	**50.31**	88.40	68.81	64.42	71.77
	PointContrast	91.73	82.41	79.89	59.82	<u>54.14</u>	48.54	87.28	69.15	63.54	70.72
	STRL	92.27	82.54	79.99	<u>61.38</u>	54.01	48.31	86.95	67.64	63.31	70.71
	ALSO	**92.57**	<u>82.88</u>	**80.24**	60.10	52.12	46.76	<u>90.71</u>	**73.94**	<u>69.21</u>	<u>72.06</u>
	E-SSL^{3D}	92.08	82.73	80.18	61.00	53.82	<u>48.58</u>	**91.15**	<u>72.68</u>	**69.32**	**72.41**

3D object detection with VoxelRCNN pre-trained on KITTI-360 and fine-tuned on KITTI 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 <u>underline</u>, respectively.

Ablation study

Spatial equivariance	Temporal equivariace	Car easy	Car moderate	Car hard	Pedestrian easy	Pedestrian moderate	Pedestrian hard	Cyclist easy	Cyclist moderate	Cyclist hard	mAP(%)
✗	✗	88.68	78.85	74.36	56.30	49.13	43.33	76.48	58.62	54.79	64.50
✗	✓	**88.98**	77.80	73.81	56.53	49.73	44.61	81.50	61.74	57.67	65.82
✓	✗	87.12	77.34	74.63	**58.66**	**50.34**	**45.19**	81.09	61.71	58.00	66.01
✓	✓	88.79	**78.93**	**75.41**	56.02	48.55	43.19	**82.85**	**64.40**	**60.53**	**66.52**

The ablation study of the spatial and temporal equivariance evaluated on the task of object detection with VoxelRCNN. The reported numbers are 3D mean average precision (%) for the "Car", "Pedestrian", and Cyclist" categories for the 3 difficulty levels and 40 recall positions.

References

[1] Xie, Saining, et al. "PointContrast: Unsupervised pre-training for 3d point cloud understanding." ECCV 2020
[2] Boulch, Alexandre, et al. "ALSO: Automotive lidar self-supervision by occupancy estimation." CVPR 2023.
[3] Huang, Siyuan, et al. "Spatio-temporal self-supervised representation learning for 3d point clouds." ICCV 2021.
[4] Jin, Zhao, et al. "Deformation and correspondence aware unsupervised synthetic-to-real scene flow estimation for point clouds." CVPR 2022.