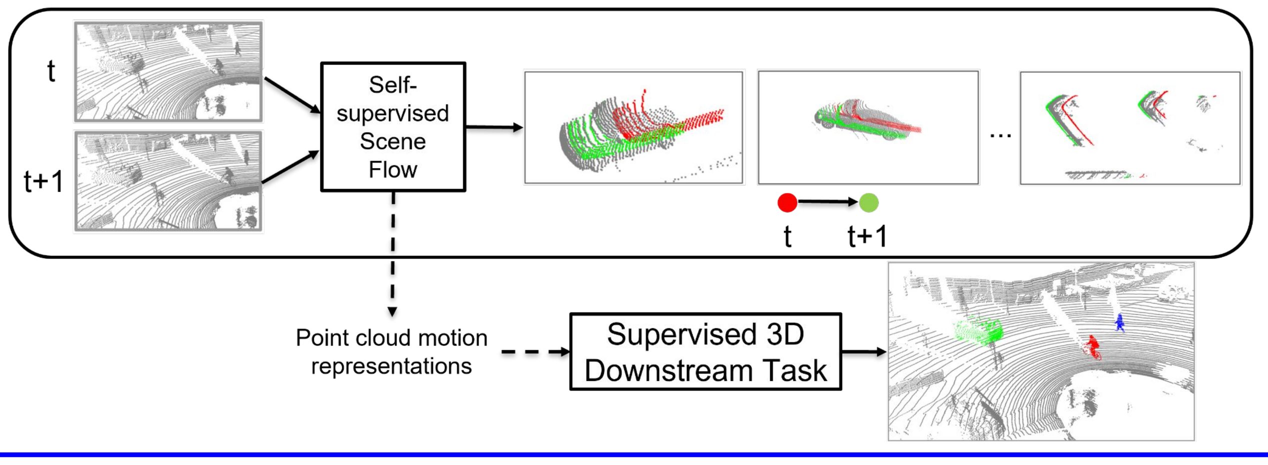
3D Object Detection with a Self-supervised Lidar Scene Flow Backbone

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]Technical University of Munich, Germany [2]Ohio State University, Columbus, USA[3] Tongji University, Shanghai, China
* Equal contribution, Contact: emec.ercelik@tum.de

ECCV TEL AVIV 2022

Motivation

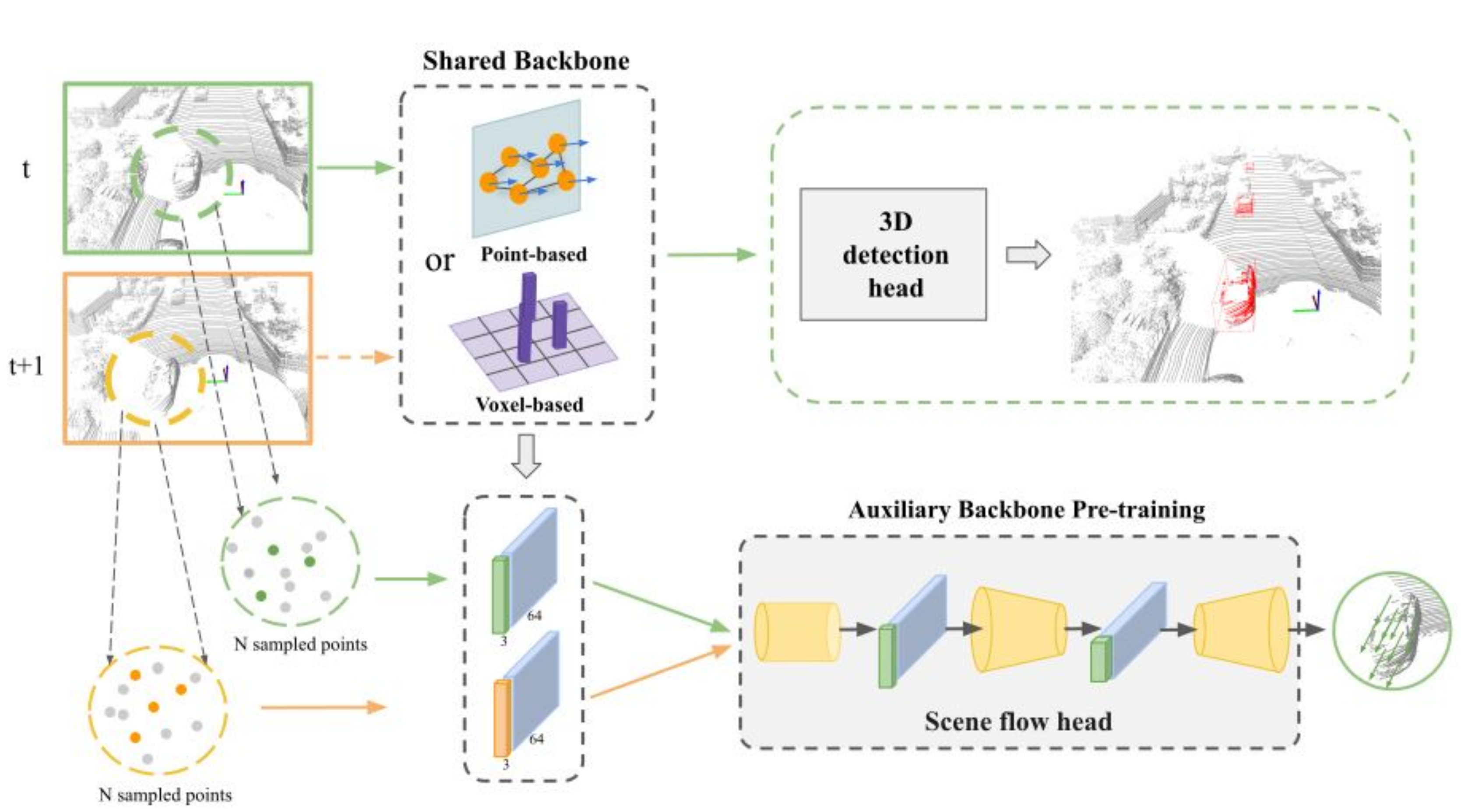
1. Self-supervised learning aims to relieve the need for **large labor-intensive labeled data** introduced by supervised learning.
2. For 3D vision tasks, self-supervised learning has been underexplored.
3. Contrary to contrastive approaches [16,17], we aim to use **inherent temporal change** in sequential lidar data by employing **self-supervised scene flow**.
4. Learned **motion representations** provide distinctive information for the 3D detector that can be used while **differentiating** objects in the environment.



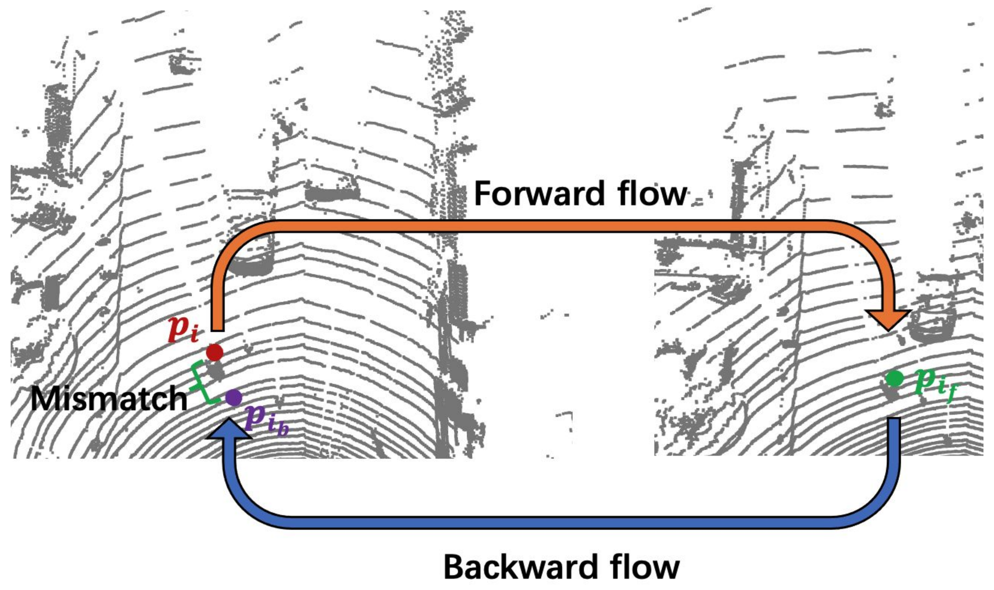
Contributions

1. Employing self-supervised point cloud scene flow estimation to learn **motion representations** for 3D object detection in tandem with supervised fine-tuning.
2. We show that auxiliary training is the best strategy for using **self-supervised cycle-consistency loss** along with supervised 3D detection loss.
3. Our strategy is **especially effective with a lesser amount of supervised data**. We obtained a significant performance boost when only a smaller part of labeled data was used for the 3D detection task.

Methodology



1. We extract features of the sampled points from **two successive frames** using the 3d detector's backbone
2. A modified **Flownet3d [5] head estimates the flow vectors** and the **self-supervised cycle consistency loss trains the head and the backbone**.
3. Then we fine-tune the **pre-trained backbone and the 3d detection head** on the smaller labelled 3d detection data.
4. We also apply **alternating training**, which repeats these two steps using trained backbone from one step prior.



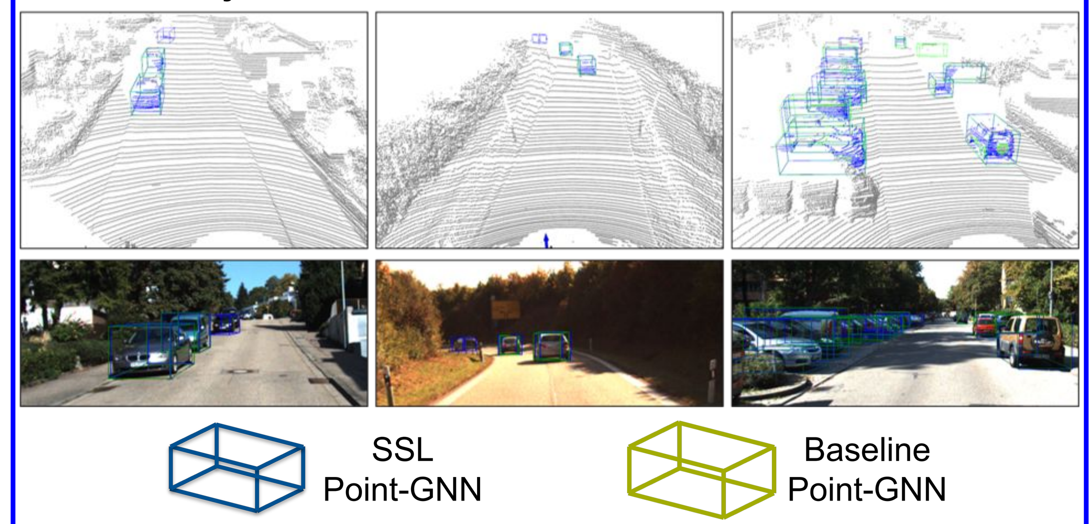
The cycle consistency loss [6] makes use of the **mismatch** of the points propagated to the same frame through the **forward and backward** passes.

Alternating training

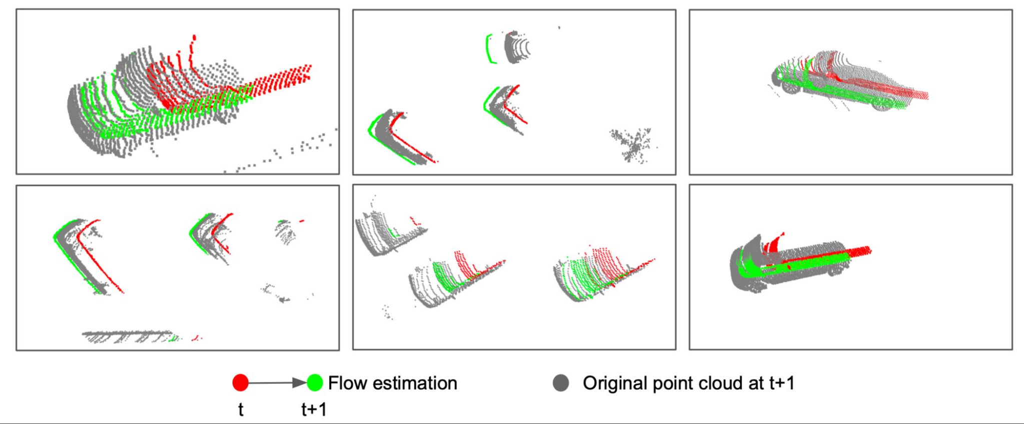
Step	Training	Backbone Init.	Head Init.
Step 1	Scene Flow	-	-
Step 2	3D detection	Step 1	-
Step 3	Scene Flow	Step 2	Step 1
Step 4	3D detection	Step 3	Step 2

Qualitative Results

3D Object Detection results on KITTI val set



Sparse scene flow on KITTI tracking



Conclusion

1. We propose a **self-supervised motion-aware backbone pre-training method for 3D object detection.**
2. **Scene flow training using the cycle consistency** helps the backbone learn distinctive features.
3. Our experimental results on nuScenes and KITTI datasets show that **our method can significantly improve 3D Detectors performance**.

Quantitative 3D Object Detection Results

Self-supervised pre-training followed by fine-tuning with all annotated data

KITTI validation set

Car (IoU=0.7)	3D $AP_{R_{40}}$			BEV $AP_{R_{40}}$		
Method	Easy	Mod	Hard	Easy	Mod	Hard
Point-GNN [3]	90.44	82.12	77.70	93.03	89.31	86.86
Self-supervised Point-GNN	**91.43**	**82.85**	**80.12**	**93.55**	**89.79**	**87.23**
Improvement	+0.99	+0.73	+2.42	+0.52	+0.48	+0.37
PointPillars[1]	85.41	73.98	67.76	89.93	86.57	85.20
Self-supervised PointPillars	**85.92**	**76.33**	**74.32**	**89.96**	**87.44**	**85.53**
Improvement	+0.51	+2.36	+6.56	+0.03	+0.87	+0.33

KITTI test set

Car (IoU=0.7)	3D $AP_{R_{40}}$			BEV $AP_{R_{40}}$		
Method	Easy	Mod	Hard	Easy	Mod	Hard
Associate-3Ddet[9]	85.99	77.40	70.53	91.40	88.09	82.96
UBER-ATG-MMF[10]	88.40	77.43	70.22	93.67	88.21	81.99
CenterNet3D[11]	86.20	77.90	73.03	91.80	88.46	83.62
SECOND[12]	87.44	79.46	73.97	92.01	88.98	83.67
SERCNN[13]	87.74	78.96	74.30	94.11	88.10	83.43
PointPillars[1]	80.51	68.57	61.79	**90.74**	84.98	79.63
Self-supervised PointPillars	**82.54**	**72.99**	**67.54**	88.92	**85.73**	**80.33**
Improvement	+2.03	+4.42	+5.75	-1.82	+0.75	+0.7

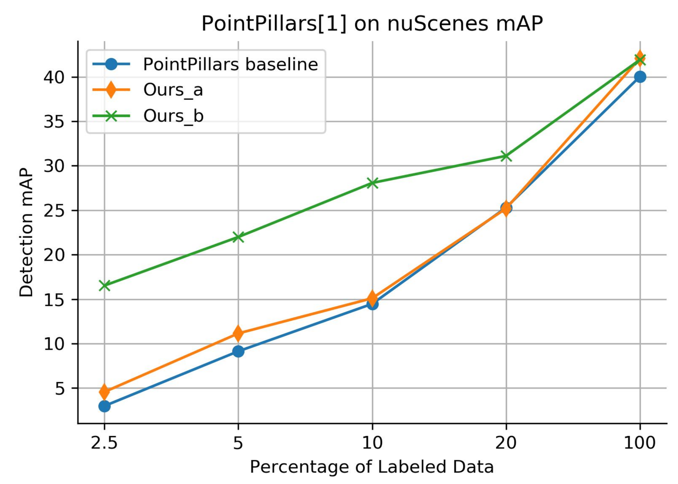
nuScenes validation set

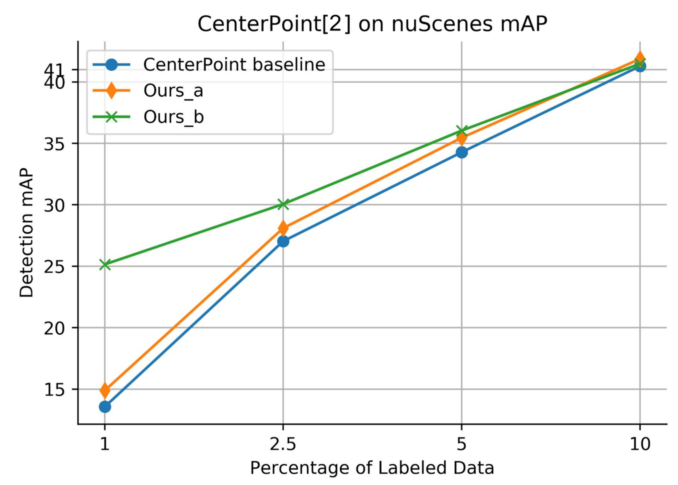
Method	mAP	NDS	Car	Ped
SECOND[12]	27.12	-	75.53	59.86
PointPillars[1]	40.02	53.29	80.60	72.40
Self-supervised PointPillars	**42.06**	**55.02**	**81.10**	**74.50**
CenterPoint[2]	49.13	59.73	83.70	77.40
Self-supervised CenterPoint	**49.94**	**60.06**	**84.10**	**77.90**

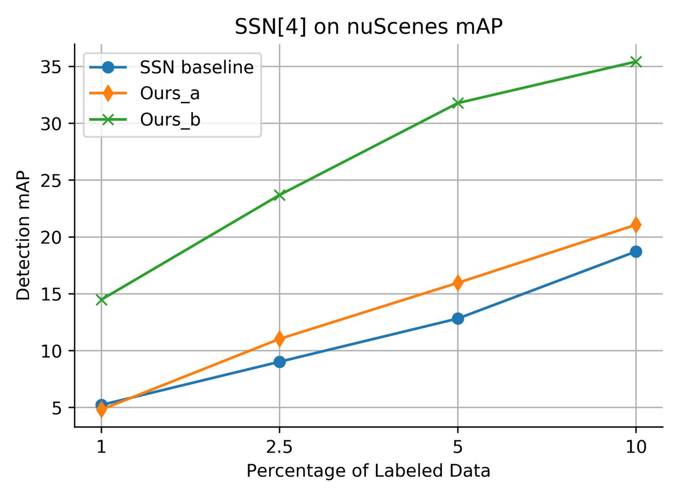
nuScenes test set

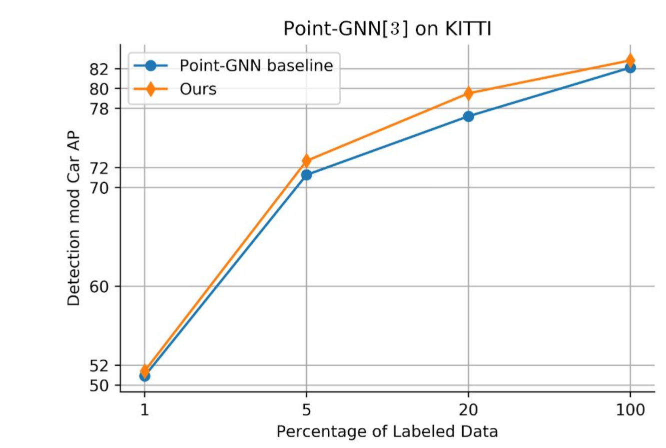
Method	mAP	NDS	Car	Ped
PointPillars[1]	30.50	45.30	68.40	59.70
InfoFocus[14]	39.50	39.50	77.90	63.40
PointPillars+[15]	40.10	55.00	76.00	64.00
Self-supervised PointPillars	**43.63**	**56.28**	**81.00**	**73.10**
CenterPoint[2]	49.54	59.64	83.40	76.10
Self-supervised CenterPoint	**51.42**	**60.92**	**83.80**	**77.00**

Supervised fine-tuning with a small set of labeled data









Comparison with other Self-supervised methods

Approach	Model	5%		10%	
		mAP	NDS	mAP	NDS
PointContrast[16]	CenterPoint[2]	30.79	41.57	38.25	50.1
GCC3D[17]		32.75	44.2	39.14	50.48
Ours		**36.04**	**48.28**	**41.29**	**51.35**

References

[1] Lang, Alex H., et al. "Pointpillars: Fast encoders for object detection from point clouds." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.
[2] Yin, Tianwei, Xingyi Zhou, and Philipp Krahenbuhl. "Center-based 3d object detection and tracking." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.
[3] Shi, Weijing, and Raj Rajkumar. "Point-gnn: Graph neural network for 3d object detection in a point cloud." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
[4] Zhu, Xinge, et al. "Ssn: Shape signature networks for multi-class object detection from point clouds." *European Conference on Computer Vision*. Springer, Cham, 2020.
[5] Liu, Xingyu, Charles R. Qi, and Leonidas J. Guibas. "Flownet3d: Learning scene flow in 3d point clouds." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.
[6] Mittal, Himangi, Brian Okorn, and David Held. "Just go with the flow: Self-supervised scene flow estimation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
[7] Geiger, Andreas, Philip Lenz, and Raquel Urtasun. "Are we ready for autonomous driving? the kitti vision benchmark suite." *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012.
[8] Caesar, Holger, et al. "nuscenes: A multimodal dataset for autonomous driving." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
[9] Du, Liang, et al. "Associate-3Ddet: Perceptual-to-conceptual association for 3D point cloud object detection." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
[10] Liang, Ming, et al. "Multi-task multi-sensor fusion for 3d object detection." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.
[11] Wang, Guojun, et al. "Centernet3d: An anchor free object detector for autonomous driving." *arXiv preprint arXiv:2007.07214* (2020).
[12] Yan, Yan, Yuxing Mao, and Bo Li. "Second: Sparsely embedded convolutional detection." *Sensors* 18.10 (2018): 3337.
[13] Zhou, Dingfu, et al. "Joint 3d instance segmentation and object detection for autonomous driving." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.
[14] Wang, Jun, et al. "Infofocus: 3d object detection for autonomous driving with dynamic information modeling." *European Conference on Computer Vision*. Springer, Cham, 2020.
[15] Vora, Sourabh, et al. "Pointpainting: Sequential fusion for 3d object detection." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
[16] Xie, Saining, et al. "Pointcontrast: Unsupervised pre-training for 3d point cloud understanding." *European conference on computer vision*. Springer, Cham, 2020.
[17] Liang, Hanxue, et al. "Exploring geometry-aware contrast and clustering harmonization for self-supervised 3D object detection." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.