Point Contrastive learning for LiDAR-based 3D object detection in autonomous driving

Efstathios Karypidis, Georgios Zamanakos, Lazaros Tsochatzidis and Ioannis Pratikakis
Department of Electrical and Computer Engineering
Democritus University of Thrace, Xanthi, 67100, Greece
Email: stathiskaripidis@gmail.com, gzamanak@ee.duth.gr, ltsochat@ee.duth.gr, ipratika@ee.duth.gr

***Abstract*—Current progress in 3D Perception tasks for autonomous driving relies upon neural network architectures that their training requires a growing demand for annotated data. However, semantic annotation of 3D scenes is a very expensive and labor-intensive task. In this paper, we present an approach for self-supervised, data-efficient learning in the context of point contrastive learning, using two distinct pre-training techniques towards improving performance in LiDAR-based 3D object detection in autonomous driving. Our experimental work relies upon standard benchmarking datasets, namely KITTI and Waymo. Under a comprehensive evaluation framework it is shown that, in the absence of large annotated data, the proposed approach could achieve improved performance.**

***Index Terms*—Self-Supervised Learning, 3D Object Detection, LiDAR, Point Clouds**

I. Introduction

Autonomous Driving (AD) is an active research field as an increasing number of car manufacturers are launching vehicles with AD capabilities in both development and commercial stages. Despite the rapid progress demonstrated in recent years, achieving autonomy remains challenging due to the complex, unpredictable and dynamic driving environment in which an autonomous vehicle operates. In order to acquire an accurate estimation of the vehicle's surroundings, AD systems employ a 'perception' pipeline that, among others, incorporates an object detection module. While camera-based 2D object detection can achieve satisfactory performance for 2D tasks, the localization of the objects in the 3D space is not a trivial task. Towards this end, more advanced sensors are being deployed, with a typical example being the LiDAR (Light Detection and Ranging) sensor, which captures a 3D point cloud, where the points correspond to the reflection of light rays emitted 360 degrees around the vehicle.

Deep neural networks (DNNs) have established state-of-the-art performance in a multitude of machine learning and computer vision tasks, including 3D object detection. Moreover, the self-supervised learning (SSL) paradigm is gaining popularity, as a means to overcome the difficulties and the cost of constructing large-scale datasets, required to support effective training of DNNs. These methods aim for the self-supervised pre-training of a general feature extractor that can be later on fine-tuned to solve a downstream task. While SSL is a highly investigated topic for natural language and image processing tasks, it has not reached a high maturity level for 3D recognition tasks. Outstanding works [1], [2] in this area have demonstrated that self-supervised pre-training on 3D indoor scenes can lead to performance boost when fine-tuned for 3D downstream tasks in other domains, with emphasis on classification and semantic segmentation. However, limited experimentation has been conducted addressing 3D object detection.

In this paper, we investigate the use of point contrastive learning on LiDAR-based 3D object detection for scenes in the context of autonomous driving. Specifically, two distinct SSL techniques, namely PointContrast and Contrastive Scene Context are used for the pre-training of a backbone network, towards extracting representations for each point in the point cloud. Consecutively, the encoder of this network is utilized as the backbone of a 3D object detector which is further fine-tuned on the available annotated data. Extensive experimentation is performed regarding different amounts of data used for fine-tuning.

II. Related Work

LiDAR-based 3D object detectors, use solely point clouds from LiDAR sensors, to predict an oriented 3D bounding box (BBox) around each detected object. Recent advances in Deep Learning along with publicly available datasets for autonomous driving, resulted in many emerged 3D object detection methods that can be mainly categorized according to their input data representation into point-based, voxel-based, projection-based and multi-representation-based methods [3], [4], [5].

Point-based methods [6] use a PointNet++ [7] 3D backbone and predict 3D objects directly from raw points. Voxel-based methods first quantize the point cloud into discrete grid representations i.e voxels as in SECOND [8] or pillars [9] and then apply 3D convolutional or PointNet [7] backbones to extract 3D features. Following, 3D features are projected into a Bird's Eye View (BEV) pseudo-image, where 2D convolutional backbones are applied to perform 3D object detection. Recently, multi-representation-based methods have emerged as in PV-RCNN [10], a two-stage detector, which utilizes a hybrid architecture that leverages both points and voxels for 3D object detection. The voxel-based representation is used in the first

979-8-3503-3959-8/23/$31.00 ©2023 IEEE

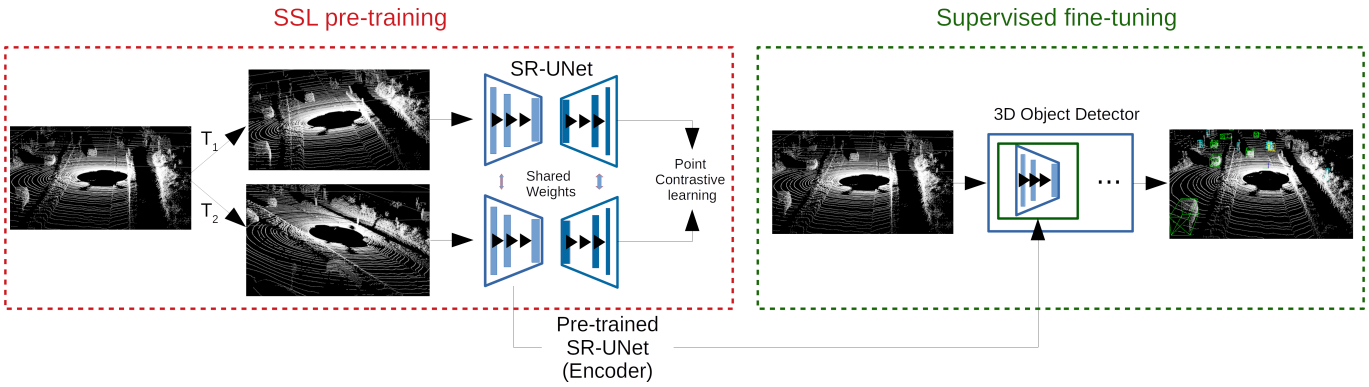

Fig. 1: The pipeline of our proposed method

stage of PV-RCNN for predicting 3D BBoxes and the point-voxel representation is used in the second stage to perform 3D BBox refinement.

Self-Supervised learning (SSL) has drawn significant attention in 2D vision tasks for improving the efficiency of learning with subsequent performance boosting in a variety of downstream tasks. However, in the case of 3D data, self-supervised learning is not widely adopted, with training from scratch on the target data still being the dominant approach.

Recently, a contrastive learning framework was proposed, namely *PointContrast* [1], which learns dense (point-level or voxel-level) representations with a focus on indoor scene point clouds. In particular, given a large unlabeled point cloud dataset, two random views that are aligned in the same world coordinates are sampled and dense correspondences between points are computed. Then, two random geometric transformations are applied in order to further transform the point clouds into two augmented views. The concept is based on learning point feature equivariance with respect to a set of random geometric transformations using an encoder/decoder architecture.

Contrastive Scene Contexts [2] extends the concept of PointContrast, by exploiting the shape contexts of a scene. For each point in a scene, a region around it is selected, given a Euclidean distance threshold. This leads to the partitioning of the scene into multiple sub-scenes, for which the contrastive loss is computed separately. The final loss is computed by averaging the contrastive loss of all sub-scenes.

III. METHODOLOGY

Current research work in point contrastive learning has examined the effectiveness of PointContrast and Contrastive Scene Contexts frameworks on downstream tasks, like indoor and outdoor classification as well as segmentation on a dense point cloud dataset for indoor scenes, namely S3DIS [11]. However, self-supervised learning for 3D object detection in outdoor scenes has not been exploited, yet.

In our work, we perform pre-training with point contrastive learning frameworks, prior to the supervised training of LiDAR-based 3D object detection in outdoor scenes used to an autonomous driving context. Self-supervised pre-training is the remedy to the lack of large annotated data by providing useful geometric priors, that could be leveraged by existing 3D object detectors towards improved performance.

In our methodology, we have examined two distinct SSL frameworks that have been used for 3D point cloud scenes, namely the *PointContrast* and *Contrastive Scene Contexts*. Both frameworks make use of a Sparse Residual U-Net (SR-UNet) [12], [13]. Pre-training based on contrastive learning is performed on outdoor scenes captured by a LiDAR sensor. For each scene, two geometric transformations (T_1, T_2) are applied to generate two views, respectively. To retain a useful representation of the scene and learn beneficial priors, the applied geometric transformations include random rotations around the roll, pitch and yaw axes following a uniform distribution of $\phi, \theta, \psi \in (-10°, 10°)$ and point jittering. The LiDAR reflection intensity attribute of the points is discarded and only the $[x, y, z]$ positions of the points are used.

For the 3D object detector networks, *SECOND* and *PV-RCNN* are chosen as the baseline networks. Once the SR-UNet [12], [13] network is pre-trained via a point contrastive learning manner, only the encoder part is kept which is further adopted by the SECOND and PV-RCNN 3D object detectors. Consecutively, fine-tuning is performed in a supervised manner, for the task of 3D object detection. Our proposed methodological pipeline is shown in Fig. 1.

IV. EXPERIMENTS

A. Datasets

KITTI [14] dataset is created by Karlsruhe Institute of Technology and Toyota Technological Institute in Chicago and it is captured with a Velodyne HDL-64E LiDAR sensor. For the task of 3D object detection, the dataset is split into a training and a test set, containing 7481 and 7518 scenes, respectively. The total number of labeled object classes is eight, however, for the task of 3D object detection, only three of them are used for evaluation, namely the 'car', 'pedestrian' and 'cyclist' classes. KITTI uses as an evaluation metric the mean Average Precision (mAP) with an Intersection over Union (IoU) threshold. **Waymo** [15] dataset contains point cloud scenes captured from 5 LiDAR sensors, installed in a single vehicle. Waymo contains training and validation labeled

TABLE I: Comparative performance of SECOND in 3D object detection, on the KITTI validation set. The results are reported in mAP with 11 recall points.

Fine-tuning Data	Pre-training Method	Car			Pedestrian			Cyclist		
		Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
2%	None	77.21	65.64	57.99	**44.21**	37.92	33.96	51.89	32.49	30.48
	PointContrast	77.24	**65.95**	**58.30**	42.72	**38.06**	**34.11**	57.20	**37.80**	**35.73**
	Contrastive Scene Contexts	**80.59**	65.36	57.70	41.83	36.89	33.25	**58.59**	35.84	33.87
5%	None	84.28	71.13	66.92	40.14	37.42	33.98	63.28	41.65	39.99
	PointContrast	83.59	70.72	66.97	42.61	38.76	35.12	**67.39**	44.77	41.70
	Contrastive Scene Contexts	**85.21**	**72.00**	**67.42**	**44.43**	**39.28**	**35.30**	65.94	**45.38**	**43.11**
10%	None	85.01	74.54	67.83	43.73	41.51	37.96	65.52	48.56	46.31
	PointContrast	85.78	**75.36**	69.66	**46.38**	**42.22**	38.01	66.20	49.60	46.98
	Contrastive Scene Contexts	**85.97**	75.13	**70.67**	45.73	41.83	**38.62**	**69.55**	**51.78**	**48.54**
20%	None	87.16	76.66	74.41	49.88	45.93	43.15	75.85	57.74	54.98
	PointContrast	87.26	76.86	**74.97**	50.45	45.78	42.64	**78.66**	**59.59**	**56.21**
	Contrastive Scene Contexts	**87.40**	**76.96**	73.77	**51.68**	**47.60**	**43.99**	76.61	58.37	55.00

TABLE II: Comparative performance of PV-RCNN in 3D object detection, on the KITTI validation set. The results are reported in mAP with 11 recall points.

Fine-tuning Data	Pre-training Method	Car			Pedestrian			Cyclist		
		Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
2%	None	76.03	63.15	57.26	38.60	35.54	**32.33**	52.33	34.59	32.41
	PointContrast	75.10	63.25	57.27	39.85	**36.26**	32.27	**58.92**	**38.27**	**35.72**
	Contrastive Scene Contexts	**76.18**	**64.24**	**57.67**	**39.95**	35.15	32.01	57.29	36.91	35.30
5%	None	87.33	76.78	71.04	**54.86**	**48.67**	43.82	72.64	52.72	49.95
	PointContrast	**87.81**	77.13	72.97	51.27	46.27	42.14	76.49	**53.64**	**51.13**
	Contrastive Scene Contexts	87.27	**77.18**	**73.04**	53.94	48.55	**43.99**	**76.79**	52.04	49.63
10%	None	88.75	78.29	76.93	49.15	45.24	41.86	80.56	**59.52**	**56.11**
	PointContrast	**88.77**	78.50	**77.40**	**57.44**	**52.12**	**48.22**	80.84	58.52	55.09
	Contrastive Scene Contexts	88.71	**78.58**	77.33	52.91	47.86	43.84	**81.35**	59.04	55.55
20%	None	**89.30**	**78.97**	**78.14**	**59.05**	**53.02**	**49.26**	**84.65**	63.68	**61.14**
	PointContrast	89.09	78.85	78.02	57.90	51.65	47.91	83.32	**63.73**	60.66
	Contrastive Scene Contexts	88.94	78.80	**78.14**	57.16	51.81	48.73	79.97	60.83	58.18

sets with 798 sequences (158.361 scenes) and 202 sequences (40.077 scenes) respectively, along with a testing set with 150 sequences. For the task of 3D object detection three classes are used, namely the 'vehicle', 'pedestrian' and 'cyclist' classes. Waymo uses as an evaluation metric the mAP.

B. Pre-training with Point Contrastive Learning

Contrastive learning training, both for PointContrast and Contrastive Scene Contexts, is performed on 80% of the KITTI training set. The networks are trained with a batch size of 8 for 2 epochs using an Adam optimizer with a learning rate equal to 10^{-3}. For PointContrast, the PointInfoNCE loss is used while for the Contrastive Scene Contexts, eight partitions are chosen to segment the scene around each point, with radius R_1, R_2 equal to 2 and 20 meters, respectively. The remainder parameters are chosen as in the official implementations of PointContrast and Contrastive Scene Contexts, respectively.

C. 3D Object Detection Training

3D Object Detection Training is performed on the KITTI training set. Out of the 7481 training scenes, 3769 scenes are used as validation samples and the rest 3712 scenes are used as training samples. The training samples are further randomly divided into groups of 2%, 5%, 10% and 20% containing 75, 186, 372 and 743 scenes, respectively.

SECOND and PV-RCNN are trained on each group of 2%, 5%, 10% and 20% of the training samples. Their performance is evaluated for the 3D object detection task on the 3769 validation scenes. The networks are trained for 80 epochs with the default training parameters and augmentations as in the OpenPCDet [16] framework. The learned weights of the 3D backbone from contrastive learning training remain unfrozen for all epochs.

D. Results

Experimental results for the performance of SECOND and PV-RCNN are shown in Table I and Table II, respectively. As it appears, for both SECOND and PV-RCNN, pre-training with point contrastive learning frameworks results in improved performance, especially when a small amount of labeled training data is used for the task of 3D object detection. This is mostly observed in the case where 2%, 5% and 10% of the data are used for training on the 3D object detection task.

For SECOND, training on the 20% of the data still results in improved overall performance for all classes and difficulty levels. However, this is not the case for PV-RCNN, as the network without any pre-training achieves a higher performance compared to its pre-trained counterparts. This can be attributed to the fact that, due to its second stage, PV-RCNN is able to learn more discriminative semantics from fewer samples, compared to the single-stage detector SECOND.

TABLE III: Ablation studies on the radius hyperparameter of Contrastive Scene Contexts. Performance is reported for SECOND in 3D object detection, on the KITTI validation set. The results are reported in mAP with 11 recall points.

Fine-tuning Data	Method	R1	R2	Car			Pedestrian			Cyclist		
				Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
2%	SECOND	2	10	75.40	64.51	57.30	**45.12**	**39.38**	**35.67**	54.93	34.63	32.89
		2	20	**80.59**	**65.36**	**57.70**	41.83	36.89	33.25	**58.59**	35.84	33.87
		4	6	76.45	64.76	57.50	44.72	38.88	34.84	56.68	35.64	33.08
		4	20	75.87	64.50	57.27	42.97	37.28	33.81	58.55	**37.12**	**34.54**

TABLE IV: Comparative performance of SECOND in 3D object detection, on the KITTI validation set. The results are reported in mAP with 11 recall points. Pre-training with Contrastive learning is performed on Waymo dataset.

Fine-tuning Data	Pre-training Method	Car			Pedestrian			Cyclist		
		Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
2%	None	**77.21**	**65.64**	**57.99**	**44.21**	37.92	33.96	51.89	32.49	30.48
	Contrastive Scene Contexts	76.81	64.77	57.37	43.29	**39.50**	**35.51**	**54.56**	**33.99**	**32.10**
5%	None	84.28	**71.13**	**66.92**	40.14	**37.42**	**33.98**	**63.28**	41.65	**39.99**
	Contrastive Scene Contexts	**84.44**	70.55	66.91	**42.38**	37.22	33.66	62.43	**42.02**	39.91
10%	None	85.01	74.54	67.83	43.73	**41.51**	**37.96**	65.52	48.56	46.31
	Contrastive Scene Contexts	**86.34**	**74.80**	**68.31**	**43.94**	40.43	37.22	**68.25**	**50.52**	**47.62**
20%	None	**87.16**	**76.66**	**74.41**	**49.88**	**45.93**	**43.15**	75.85	57.74	54.98
	Contrastive Scene Contexts	86.55	76.64	73.25	46.66	42.91	40.11	**76.61**	**58.65**	**55.30**

E. Ablation Study

1) Local over Global Scene Contexts: For Contrastive Scene Contexts, the hyperparameter values R_1, R_2 define the receptive field of the area under which the shape contexts of a scene are represented. Table III demonstrates the performance of SECOND, for different R_1, R_2 values. A smaller receptive field, R_1, R_2 equal to 2 and 10 results in increased performance for pedestrians, followed by the receptive field of R_1, R_2 equal to 4 and 6. This demonstrates the importance of local over global contexts for pedestrian detection. For cyclists, an increased performance is demonstrated for a larger receptive field, when R_1, R_2 equal to 4 and 20, along with R_1, R_2 equal to 2 and 20. Despite being an object of relatively small spatial size, for cyclist detection, the global contexts appear to be more important than local contexts. For cars, it appears that a balance between local and global contexts, when R_1, R_2 equal to 2 and 20, results in the best detection results.

2) SSL from other domain: For Contrastive Scene Contexts, the contrastive learning pre-training is also performed in another domain for autonomous driving, namely Waymo. The SR-UNet is trained for 1 epoch in Waymo dataset. Table IV shows the performance for SECOND in 3D object detection for KITTI, having as a prior the contrastive learning from Waymo. As it appears, the performance of SECOND is improved, especially in cases where few labeled scenes are available. This demonstrates the potential of learning useful geometric priors from other domains and raises a research interest for future work.

V. CONCLUSIONS

In this paper, we investigated the use of point contrastive learning frameworks, by learning geometric priors that are used for the task of 3D object detection in autonomous driving. We performed extensive experiments by adapting and integrating two contrastive learning frameworks for self-supervised pre-training on LiDAR point clouds. We evaluated the effectiveness of the pre-trained geometric priors by incorporating them into two 3D object detector networks and evaluated their performance for the task of 3D object detection in KITTI dataset. Experimental results demonstrate that pre-training with contrastive learning can increase the performance of 3D object detectors, especially when a limited number of labeled scenes are available. Additionally, we have performed ablation studies for identifying the importance of local and global contexts for each object class, along with the potential of learning beneficial geometric priors from other domains.

ACKNOWLEDGMENT

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH-CREATE-INNOVATE (project code:T2EDK-02743). We would also like to thank NVIDIA Corporation, which kindly donated the Titan X GPU, that has been used for this research.

REFERENCES

[1] S. Xie, J. Gu, D. Guo, C. R. Qi, L. Guibas, and O. Litany, "PointContrast: Unsupervised pre-training for 3D point cloud understanding," in *European conference on computer vision*. Springer, 2020, pp. 574–591.

[2] J. Hou, B. Graham, M. Nießner, and S. Xie, "Exploring data-efficient 3D scene understanding with contrastive scene contexts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 587–15 597.

[3] R. Qian, X. Lai, and X. Li, "3D object detection for autonomous driving: A survey," *Pattern Recognition*, vol. 130, p. 108796, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320322002771

[4] G. Zamanakos, L. Tsochatzidis, A. Amanatiadis, and I. Pratikakis, "A comprehensive survey of lidar-based 3D object detection methods with deep learning for autonomous driving," *Computers & Graphics*, vol. 99, pp. 153–181, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0097849321001321

[5] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, "Deep learning for 3D point clouds: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 12, pp. 4338–4364, 2021.

[6] S. Shi, X. Wang, and H. Li, "PointRCNN: 3D object proposal generation and detection from point cloud," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 770–779.

[7] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 652–660.

[8] Y. Yan, Y. Mao, and B. Li, "SECOND: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[9] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 689–12 697.

[10] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "PV-RCNN: Point-voxel feature set abstraction for 3D object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 529–10 538.

[11] I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese, "3D semantic parsing of large-scale indoor spaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1534–1543.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[13] C. Choy, J. Gwak, and S. Savarese, "4D spatio-temporal convnets: Minkowski convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3075–3084.

[14] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.

[15] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2446–2454.

[16] O. D. Team, "OpenPCDet: An open-source toolbox for 3D object detection from point clouds," https://github.com/open-mmlab/OpenPCDet, 2020.