T-MAE: Temporal Masked Autoencoders for Point Cloud Representation Learning

Weijie Wei, Fatemeh Karimi Nejadasl, Theo Gevers, and Martin R. Oswald

University of Amsterdam, the Netherlands

Abstract. Temporal information, which is inherent in the LiDAR point cloud sequence, is consistently disregarded in point cloud representation learning. To better utilize this property, we propose an effective pre-training strategy, namely Temporal Masked Auto-Encoders (T-MAE), which takes as input temporally adjacent frames and learns temporal dependency. A SiamWCA backbone, containing a Siamese encoder and a windowed cross-attention (WCA) module, is established for the two-frame input. SiamWCA is a powerful architecture but heavily relies on annotated data. Our T-MAE pre-training strategy alleviates its demand for annotated data. Comprehensive experiments demonstrate that T-MAE achieves the best performance on both Waymo and ONCE datasets among competitive self-supervised approaches.

Keywords: Self-supervised learning · LiDAR point cloud · 3D detection

1 Introduction

Many self-supervised learning (SSL) methods for point clouds understanding in autonomous driving rely on contrastive learning [9–11, 16, 20]. These approaches model the similarity and dissimilarity between entities, such as segments [10, 15] and/or points [16]. In the wake of masked image modeling as a pretext task [4], efforts have also been devoted to the reconstruction of masked points [8,14,17,19]. The main idea is randomly masking points or voxels and urging the network to infer the coordinates of points [19] and/or voxels [17] or other properties, *e.g.*, occupancy [1,8] and curvature [14]. Nevertheless, these methodologies often operate within the confines of a single-frame scenario, disregarding the fact that LiDAR data is typically acquired on a frame-by-frame basis. In other words, the valuable semantic information in temporally adjacent frames is barely exploited.

Several methods attempt to leverage temporal information [5, 6, 10, 15] by incorporating multi-frame input during the self-supervised phase but their core concepts remain grounded in contrastive learning. Specifically, the point clouds captured at different times are treated as augmented samples of the same scene, without including temporal correspondence into the modeling procedure.

Therefore, we propose a new SSL paradigm, namely T-MAE, to exploit the accumulated observations. Our **contributions** are summarized as follows: **1)** We propose T-MAE, a novel and effective SSL approach for representation learning

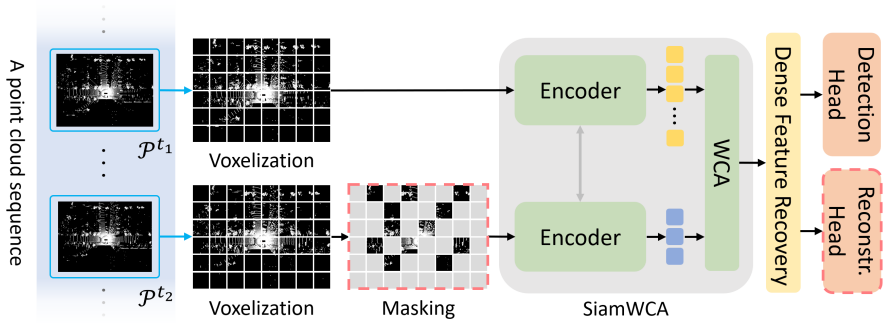

Fig. 1: Overview of our architecture and the proposed T-MAE pre-training. Two frames are sampled from a sequence of point clouds and are voxelized. During pre-training, the current frame $\mathcal{P}^{t_2}$ undergoes an additional masking process. Note that the dashed boxes indicate operations for pre-training phase only. Next, voxel-wise tokens are computed by a Siamese encoder. The two-way gray arrow indicates weight sharing. The WCA module takes as input the full tokens of the previous frame and the partial observation of the current frame and outputs enhanced tokens. The dense feature recovery places sparse tokens back to a dense feature map and convolves the map to fill empty locations. Subsequently, the feature map is either fed to a reconstruction head that recovers masked points, or to a detection head predicting bounding boxes.

of sparse point clouds, that learns temporal modeling in the process of reconstructing masked points. **2)** We design a SiamWCA backbone to incorporate historical information. **3)** Our experiments demonstrate the efficacy of T-MAE by attaining substantial improvements on the Waymo and ONCE datasets.

2 Method

The goal of our framework is to learn a powerful representation for point clouds and integrate features from the past for present use. Therefore, two frames are sampled from a sequence of point clouds and denoted as $\mathcal{P}^{t_1}$ and $\mathcal{P}^{t_2}$, as shown in Fig. 1. During the **T-MAE** pre-training stage, the current scan is voxelized with a high masking ratio, while the previous scan is fed entirely to the encoder. Then, the pretext task is to reconstruct the current scan by incorporating voxel embeddings of the past scan, visible voxel embeddings of the current scan, and the position of masked voxels. This way, the proposed windowed cross-attention module learns to incorporate historical information into the current frame using unlabeled data. The T-MAE pre-training strategy endows the network with both a powerful representation for sparse point clouds and the capacity to strengthen the present by learning from the past.

SiamWCA. It compromises a Siamese encoder and a windowed cross-attention module. A Siamese encoder [2] is a two-branch network where both branches share the same configurations and weights. In this work, it is utilized to encode the pillar-wise representations of both frames to sparse tokens. These tokens serve as input to the WCA module which facilitates the interaction between

Table 1: Comparison with SSL methods on the Waymo validation set [12]. Random initialization denotes training from scratch. † represents duplicating the current frame as input during inference. Results for MV-JAR [17] are taken from the original paper. * and ** indicate reproduced by us and taken from GD-MAE [19], respectively. Best results are highlighted as **first**, second, and third. Differences between T-MAE pre-training and random initialization are highlighted in red.

Data Amount	Initialization	Overall		Vehicle		Pedestrian		Cyclist	
		mAP	mAPH	mAP	mAPH	mAP	mAPH	mAP	mAPH
5%	Random	43.68	40.29	54.05	53.50	53.45	44.76	23.54	22.61
	MV-JAR [17]	50.52	46.68	56.47	56.01	57.65	47.69	37.44	36.33
	GD-MAE [19]*	48.23	44.56	56.34	55.76	55.62	46.22	32.72	31.69
	T-MAE†	50.89	47.22	57.06	56.05	58.95	52.62	36.64	32.99
	T-MAE (Ours)	**51.47**$^{\uparrow 7.79}$	**49.46**$^{\uparrow 9.17}$	**57.13**	**56.63**	**59.69**	**55.28**	**37.61**	**36.48**
10%	Random	56.05	53.13	59.78	59.27	60.08	53.04	48.28	47.08
	MV-JAR [17]	57.44	54.06	58.43	58.00	63.28	54.66	50.63	49.52
	GD-MAE [19]*	57.67	54.31	59.72	59.19	60.43	52.21	52.85	51.52
	T-MAE†	58.52	55.59	60.26	59.75	62.89	55.85	52.43	51.16
	T-MAE (Ours)	**59.93**$^{\uparrow 3.88}$	**57.99**$^{\uparrow 4.86}$	**60.27**	**59.77**	**65.23**	**61.10**	**54.29**	**53.09**
20%	Random	60.21	57.61	61.58	61.08	64.63	58.41	54.42	53.33
	MV-JAR [17]	62.28	59.15	61.88	61.45	66.98	59.02	57.98	57.00
	GD-MAE [19]*	62.32	59.09	62.27	61.79	66.12	58.06	58.57	57.42
	T-MAE†	62.37	60.17	62.19	61.72	67.18	62.18	57.74	56.59
	T-MAE (Ours)	**63.52**$^{\uparrow 3.31}$	**61.80**$^{\uparrow 4.19}$	**63.10**	**62.59**	**68.23**	**64.66**	**59.23**	**58.15**
100%	Random	71.30	69.13	69.05	68.62	73.77	68.80	71.09	69.97
	MV-JAR [17]	69.16	66.20	65.52	65.12	72.77	65.28	69.19	68.20
	GD-MAE [19]**	70.62	67.64	68.72	68.29	72.84	65.47	70.30	69.16
	T-MAE†	71.56	69.00	69.39	68.95	74.42	68.43	70.86	69.61
	T-MAE (Ours)	**72.30**$^{\uparrow 1.00}$	**70.52**$^{\uparrow 1.39}$	**69.34**	**68.89**	**75.79**	**72.01**	**71.78**	**70.65**

historical and current tokens. WCA is essentially a group of cross-attention layers where the query comes from $\mathcal{P}^{t_2}$ and the key-value comes from $\mathcal{P}^{t_1}$. Note that the WCA module divides the 3D space into non-overlapping windows and then performs cross-attention within the windows for efficient computing.

3 Experiments

We present the comparison with SOTA methods on the Waymo Open dataset [12] and ONCE dataset [7].

The impact of the T-MAE pre-training. The bottom block in Tab. 1 shows that the randomly initialized SiamWCA (denoted as *Random*) performs better than any other models that are initialized with a different pre-training strategy (*i.e.* 69.13 *v.s.* 67.64), suggesting that SiamWCA is a powerful backbone capable of learning temporal modeling when provided with sufficient annotated data. However, its performance drops significantly when finetuning data is limited (*e.g.*, 40.29 *v.s.* 46.68 at 5% level), indicating a strong demand for annotated data. As a comparison, the proposed T-MAE consistently enhances SiamWCA compared to random initialization. Moreover, as the labeled data shrinks from 100% to 5%, the impact of the T-MAE pre-training becomes more pronounced, namely increasing from 1.39 to 9.17, suggesting that the pre-training approach learns a powerful representation and alleviates the demand for annotated data.

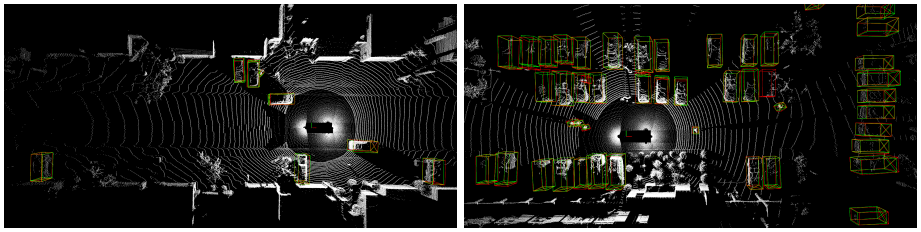

Fig. 2: Qualitative results. We depict ground truth and predictions as boxes colored in red and green for two exemplary scenes from the Waymo dataset [12].

Table 2: Performance comparisons on the validation split of the ONCE dataset [7]. Pt. indicates the model is initialized with pre-trained weights. Results for other methods are taken from GD-MAE [19].

Methods	Pt.	mAP	Vehicle				Pedestrian				Cyclist			
			Overall	0-30m	30-50m	50m-Inf	Overall	0-30m	30-50m	50m-Inf	Overall	0-30m	30-50m	50m-Inf
SECOND [18]	✗	51.89	71.19	84.04	63.02	47.25	26.44	29.33	24.05	18.05	58.04	69.96	52.43	34.61
w/ BYOL [3]	✓	51.63	71.32	83.59	64.89	50.27	25.02	27.06	22.96	17.04	58.56	70.18	52.74	36.32
w/ PointContrast [16]	✓	53.59$^{\uparrow 1.70}$	71.87	86.93	62.85	48.65	28.03	33.07	25.91	14.44	60.88	71.12	55.77	36.78
w/ DeepCluster [13]	✓	53.72$^{\uparrow 1.83}$	72.89	83.52	67.09	50.38	30.32	34.76	26.43	18.33	57.94	69.18	52.42	34.36
SPT [19]	✗	62.62	75.64	87.21	70.10	53.21	45.92	54.78	37.84	22.56	66.30	78.12	60.52	42.05
w/ GD-MAE [19]	✓	64.92$^{\uparrow 2.30}$	76.79	88.01	71.70	55.60	48.84	58.70	37.30	**25.72**	69.14	80.29	64.58	45.14
SiamWCA (Ours)	✗	63.71	76.47	87.63	71.59	55.16	47.27	57.57	36.99	21.79	67.40	78.39	62.78	43.90
w/ T-MAE (Ours)	✓	**67.00**$^{\uparrow 3.29}$	**78.35**	**88.45**	**73.05**	**57.16**	**52.57**	**62.66**	**44.18**	25.29	**70.09**	**81.14**	**65.33**	**46.48**

Comparison with SOTA methods. We aim to leverage the temporal information between two adjacent frames, which is often overlooked by other methods. This absence makes it challenging to compare our method with others in the same setting. Therefore, we implement a test-time single-frame baseline (denoted as T-MAE†) by replicating the same frame and inputting them into our pre-trained model during evaluation. Table 1 shows that T-MAE with identical frames outperforms SOTA counterparts. Moreover, T-MAE with adjacent frames achieves new SOTA at all levels in terms of overall and class-specific metrics. Figure 2 shows two exemplary qualitative results from the Waymo dataset.

To assess the generalization capabilities of our method, we also conducted experiments on the ONCE dataset [7]. As shown in Tab. 2, T-MAE outperforms other methods in most metrics, indicating its superiority. Moreover, the substantial improvement for pedestrians generalizes to this new dataset, indicating the dominance of T-MAE in pedestrian detection.

4 Conclusion

We introduced Temporal Masked Autoencoders (T-MAE), a novel self-supervised paradigm for LiDAR point cloud pre-training. Building upon the single-frame MAE baseline, we incorporated historical frames into the representation using SiamWCA, with the proposed WCA module playing a pivotal role. This pre-training enabled the model to acquire robust representations and the ability to capture motion even with very limited labeled data. Our experiments on the Waymo dataset and the ONCE dataset demonstrate the effectiveness of our approach by showing improvements over state-of-the-art methods.

References

1. Boulch, A., Sautier, C., Michele, B., Puy, G., Marlet, R.: Also: Automotive lidar self-supervision by occupancy estimation. In: CVPR (2023)
2. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a "siamese" time delay neural network. In: NeurIPS. p. 737–744 (1993)
3. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent: A new approach to self-supervised learning. In: NeurIPS (2020)
4. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2021)
5. Huang, S., Xie, Y., Zhu, S.C., Zhu, Y.: Spatio-temporal self-supervised representation learning for 3d point clouds. In: ICCV (2021)
6. Liang, H., Jiang, C., Feng, D., Chen, X., Xu, H., Liang, X., Zhang, W., Li, Z., Van Gool, L.: Exploring geometry-aware contrast and clustering harmonization for self-supervised 3d object detection. In: ICCV. pp. 3273–3282 (2021)
7. Mao, J., Niu, M., Jiang, C., Liang, H., Chen, J., Liang, X., Li, Y., Ye, C., Zhang, W., Li, Z., Yu, J., Xu, H., Xu, C.: One million scenes for autonomous driving: Once dataset. In: NeurIPS (2021)
8. Min, C., Xu, X., Zhao, D., Xiao, L., Nie, Y., Dai, B.: Occupancy-MAE: Self-supervised pre-training large-scale lidar point clouds with masked occupancy autoencoders. IEEE Transaction on Intelligent Vehicles (2022)
9. Nunes, L., Marcuzzi, R., Chen, X., Behley, J., Stachniss, C.: SegContrast: 3D point cloud feature representation learning through self-supervised segment discrimination. IEEE Robotics and Automation Letters (RA-L) **7**(2), 2116–2123 (2022)
10. Nunes, L., Wiesmann, L., Marcuzzi, R., Chen, X., Behley, J., Stachniss, C.: Temporal consistent 3d LiDAR representation learning for semantic perception in autonomous driving. In: CVPR (2023)
11. Pang, B., Xia, H., Lu, C.: Unsupervised 3d point cloud representation learning by triangle constrained contrast for autonomous driving. In: CVPR (2023)
12. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: CVPR. pp. 2443–2451 (2020)
13. Tian, K., Zhou, S., Guan, J.: Deepcluster: A general clustering framework based on deep learning. In: Machine Learning and Knowledge Discovery in Databases. pp. 809–825 (2017)
14. Tian, X., Ran, H., Wang, Y., Zhao, H.: GeoMAE: Masked geometric target prediction for self-supervised point cloud pre-training. In: CVPR (2023)
15. Wu, Y., Zhang, T., Ke, W., Susstrunk, S., Salzmann, M.: Spatiotemporal self-supervised learning for point clouds in the wild. In: CVPR. pp. 5251–5260 (2023)
16. Xie, S., Gu, J., Guo, D., Qi, C.R., Guibas, L.J., Litany, O.: Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In: ECCV (2020)
17. Xu, R., Wang, T., Zhang, W., Chen, R., Cao, J., Pang, J., Lin, D.: MV-JAR: Masked voxel jigsaw and reconstruction for LiDAR-based self-supervised pre-training. In: CVPR (2023)
18. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors **18**(10), 3337 (2018)

19. Yang, H., He, T., Liu, J., Chen, H., Wu, B., Lin, B., He, X., Ouyang, W.: GD-MAE: Generative decoder for MAE pre-training on LiDAR point clouds. In: CVPR (2023)
20. Yin, J., Zhou, D., Zhang, L., Fang, J., Xu, C.Z., Shen, J., Wang, W.: Proposalcontrast: Unsupervised pre-training for lidar-based 3d object detection. In: ECCV (2022)