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

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

Universiteit van Amsterdam · computer vision lab · ATLAS LAB.

ECCV MILANO 2024

Introduction

Motivation

- In autonomous driving, LiDAR points are dynamic **temporal sequence**.
- It is essential to integrate **historical observation** for present decision-making.

Existing Self-Supervised Learning (SSL)

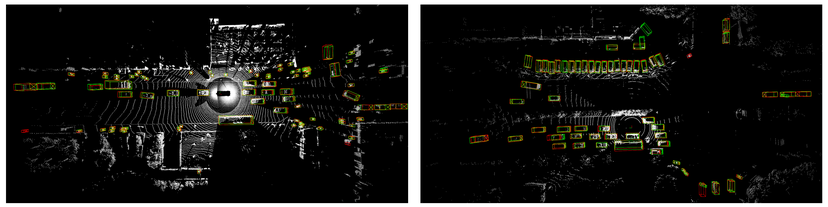
BYOL, DINO, MAE ➔ One frame, ignoring temporal info!

Temporal contrastive learning ➔ Use multiple frames as data augmentations, disregarding temporal correspondence.

Results

ONCE Detection Dataset

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
PV-RCNN [48]	✗	53.55	77.77	**89.39**	72.55	**58.64**	23.50	25.61	22.84	17.27	59.37	71.66	52.58	36.17
IA-SSD [73]	✗	57.43	70.30	83.01	62.84	47.01	39.82	47.45	32.75	18.99	62.17	73.78	56.31	39.53
CenterPoint-Pillar [68]	✗	59.07	74.10	85.23	69.22	53.14	40.94	48.43	34.72	20.09	62.17	73.70	56.05	40.19
CenterPoint-Voxel [68]	✗	60.05	66.79	80.10	59.55	43.39	49.90	56.24	42.61	**26.27**	63.45	74.28	57.94	41.48
SECOND [63]	✗	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 [20]	✓	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 [59]	✓	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 [52]	✓	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 [64]	✗	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 [64]	✓	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**



* Green box: ground-truth. Red box: prediction.

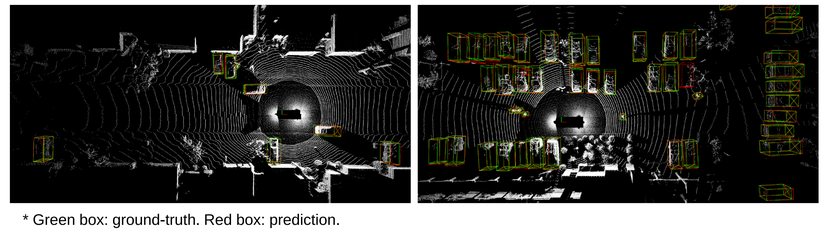
Results

Waymo Detection Dataset

- SOTA performance at all data level
- Significant boost when annotation is limited
- Notable improvement for *Pedestrian* class

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
	PointContrast [59]	45.32	41.30	52.12	51.61	53.68	43.22	30.16	29.09
	ProposalContrast [67]	46.62	42.58	52.67	52.19	54.31	43.82	32.87	31.72
	MV-JAR [60]	50.52	46.68	56.47	56.01	57.65	47.69	37.44	36.33
	GD-MAE [64]*	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
	PointContrast [59]	53.69	49.94	54.76	54.30	59.75	50.12	46.57	45.39
	ProposalContrast [67]	53.89	50.13	55.18	54.71	60.01	50.39	46.48	45.28
	MV-JAR [60]	57.44	54.06	58.43	58.00	63.28	54.66	50.63	49.52
	GD-MAE [64]*	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
	PointContrast [59]	59.35	55.78	58.64	58.18	64.39	55.43	55.02	53.73
	ProposalContrast [67]	59.52	55.91	58.69	58.22	64.53	55.45	55.36	54.07
	MV-JAR [60]	62.28	59.15	61.88	61.45	66.98	59.02	57.98	57.00
	GD-MAE [64]*	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
	GCC-3D [32]	65.29	62.79	63.97	63.47	64.23	58.47	67.88	66.44
	BEV-MAE [33]	66.92	64.45	64.78	64.29	66.25	60.53	69.73	68.52
	PointContrast [59]	68.06	64.84	64.24	63.82	71.92	63.81	68.03	66.89
	ProposalContrast [67]	68.17	65.01	64.42	64.00	71.94	63.94	68.16	67.10
	MV-JAR [60]	69.16	66.20	65.52	65.12	72.77	65.28	69.19	68.20
	GD-MAE [64]**	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**

Each block indicates finetuning with x% of annotated data. *Random* denotes training from scratch. Best results are highlighted as **first**, second, and third. Differences between T-MAE pre-training and random initialization are highlighted in red.



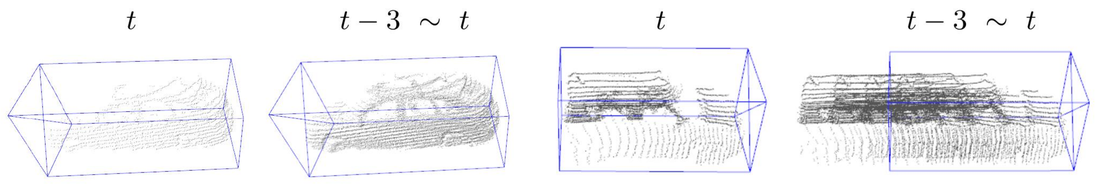
* Green box: ground-truth. Red box: prediction.

Methodology

Our Solution

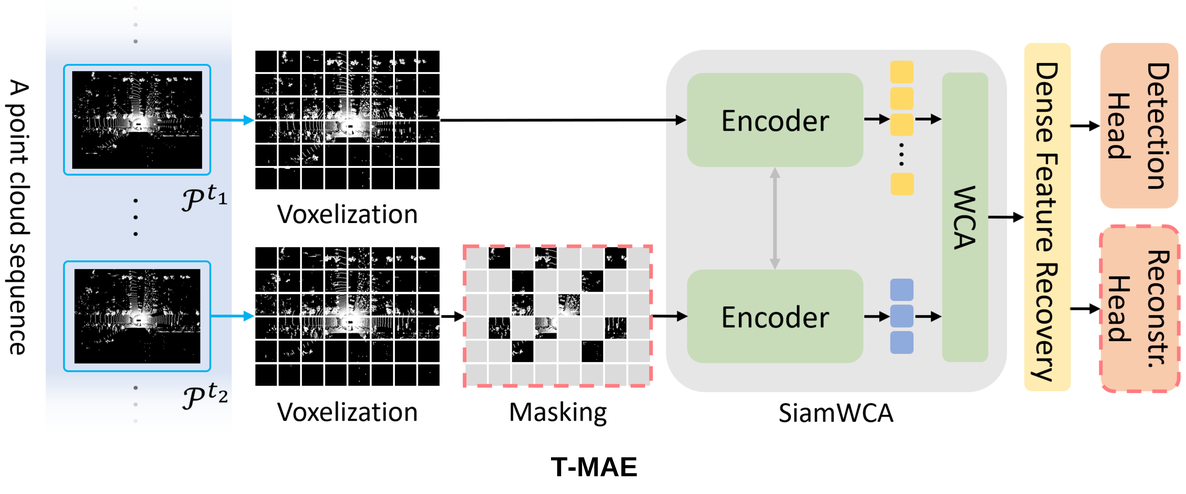
Integrate historical observation for current prediction.

- Impact of moving objects 👉
  - Fusion in latent space.
  - **Sparse Windowed Cross-Attention** (WCA).
- No enough annotated data
  - Self-supervised learning.
  - Pretext task: **Reconstruct the current frame by observing one previous frame.**
  - Inherit the pretrained weights of both the backbone and WCA.
  - Random temporal gap improves robustness.



(a) A stationary vehicle. (b) A moving vehicle.

Issue of integrating multiple frame. Integrating consecutive frames enhances stationary objects but introduces spurious points and deface moving objects.



Ablation Study

- Pretrained weights of both the Siamese encoder and WCA boost the performance.

Initialization	Pre-trained		Overall		Vehicle		Pedestrian		Cyclist	
	SE	WCA	mAP	mAPH	mAP	mAPH	mAP	mAPH	mAP	mAPH
Random	✗	✗	43.68	40.29	54.05	53.50	53.45	44.76	23.54	22.61
Partially random	✓	✗	48.19$^{\uparrow 4.51}$	45.16$^{\uparrow 4.87}$	55.91	55.38	56.29	48.74	32.38	31.36
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**