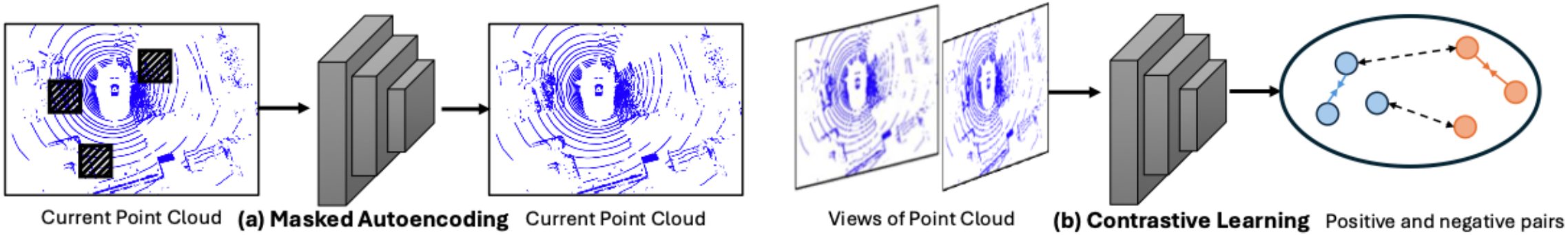
TREND: Unsupervised 3D Representation Learning via Temporal Forecasting for LiDAR Perception

Runjian Chen, Hyoungseob Park, Bo Zhang, Wenqi Shao, Ping Luo*, Alex Wong*

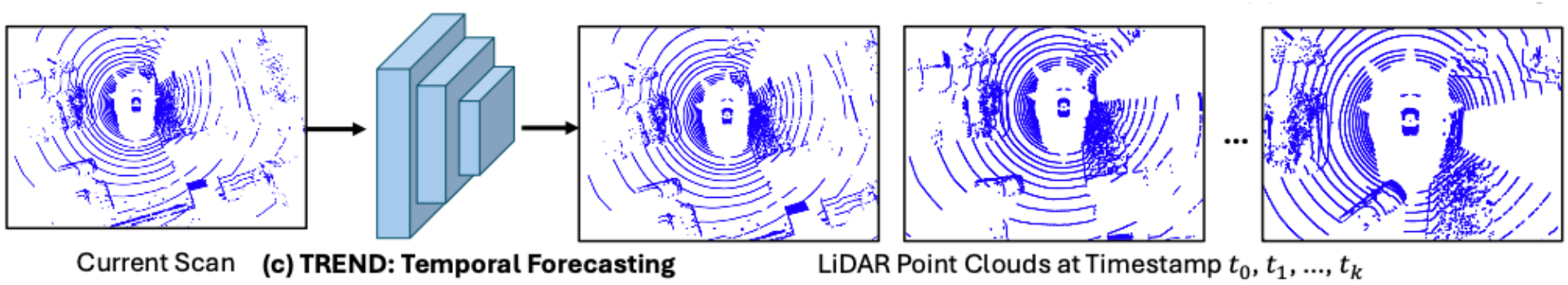
Motivation

- Previous unsupervised 3D representation learning methods including masked autoencoding and contrastive learning, rely on a predefined set of nuisance variability and pre-train the 3D backbone to be invariant to the nuisance variability. For MAE-based methods, the set of nuisance variability is occlusions and it is handcrafted set of LiDAR transformation for contrastive learning. While the procedures are unsupervised, they implicitly select the set of invariants, which benefits the downstream tasks.



Overview of TREND

- Unlike previous methods, we subscribe to allowing the 3D backbone to determine nuisances by simply observing and predicting scene dynamics. This leads to TREND, a novel unsupervised 3D representation learning approach based on forecasting LiDAR point clouds. Naturally, points belonging to the same object instance tend to move together. By observing current point cloud and predicting future observation, our pre-training scheme implicitly encodes semantics and biases of object interactions over time.



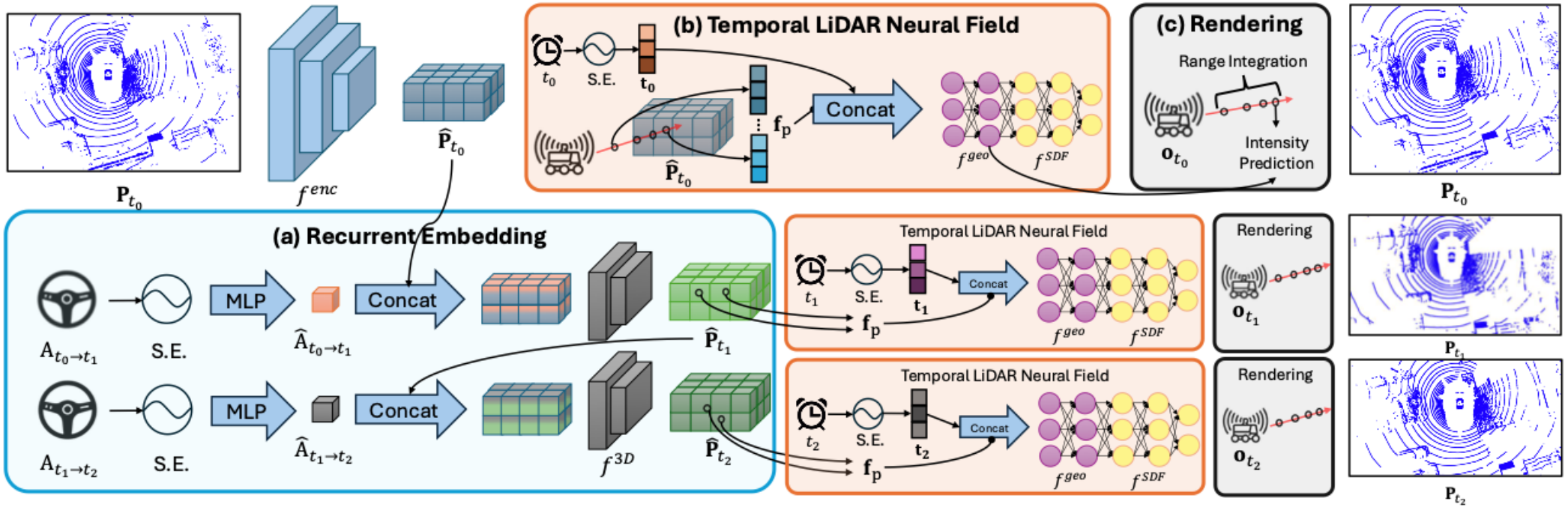
Challenges

- It is non-trivial to incorporate temporal forecasting for unsupervised 3D representation learning.
- Two main challegenes:

- How to generate 3D embeddings at different timesteps with current LiDAR scan?
- How to represent the 3D scene with the temporal 3D embeddings and optimize the network via forecasting?

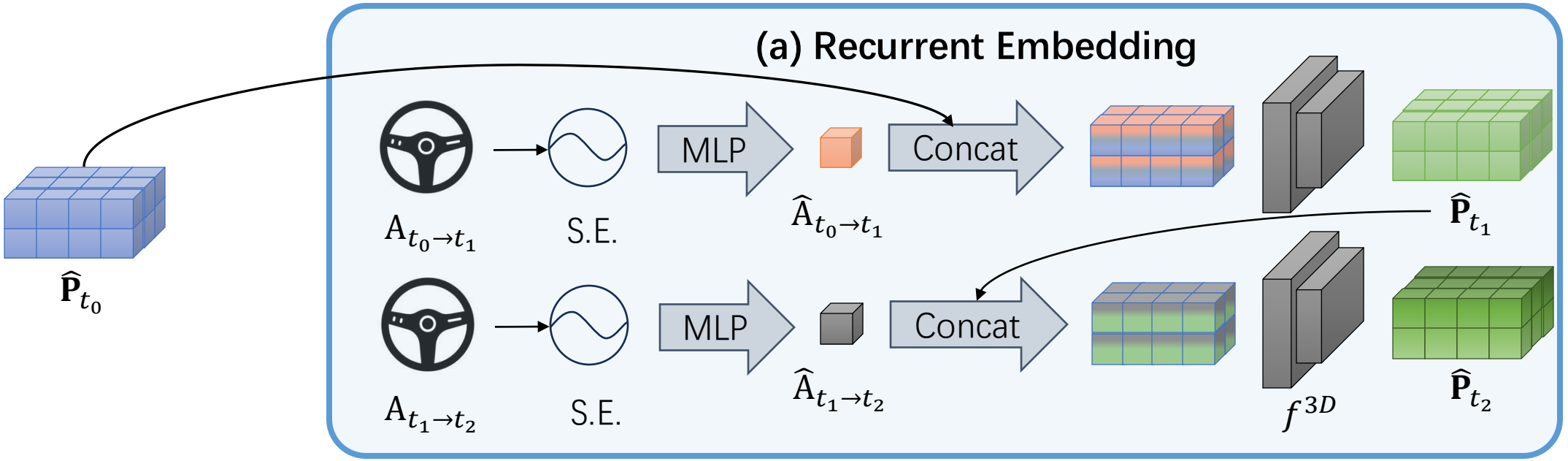
Method Overview

- **TREND** is consisted of three main parts: (a) Recurrent Embedding, (b) Temporal LiDAR Neural Field, (c) Neural rendering for both reconstruction and forecasting.



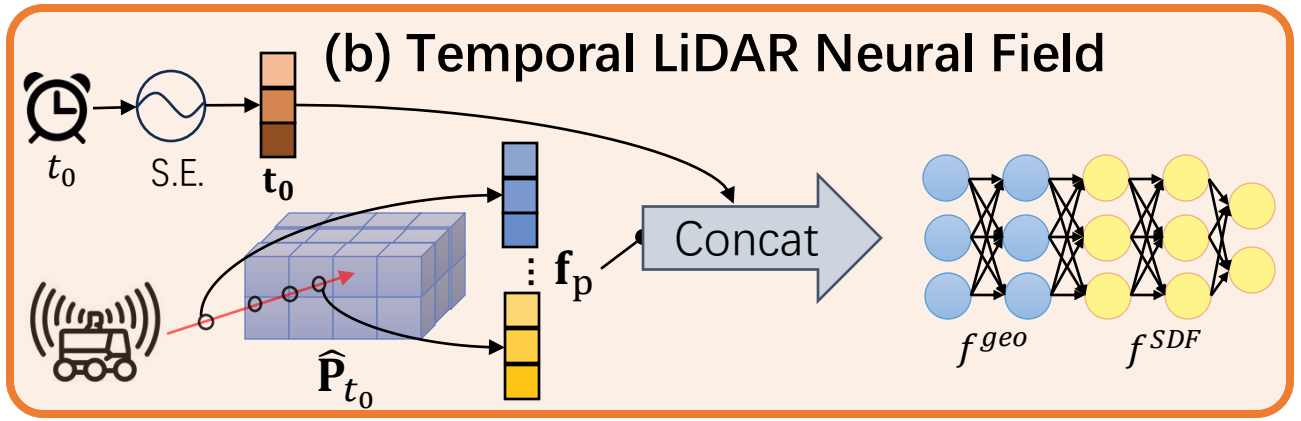
Method: Recurrent Embedding

- **Ego-action aware temporal embedding:** as the action of ego-vehicle reflects the interaction between the ego-vehicle and other traffic participants, we incorporate the ego-action to evolve the feature at timestep t_0 to future timestep.



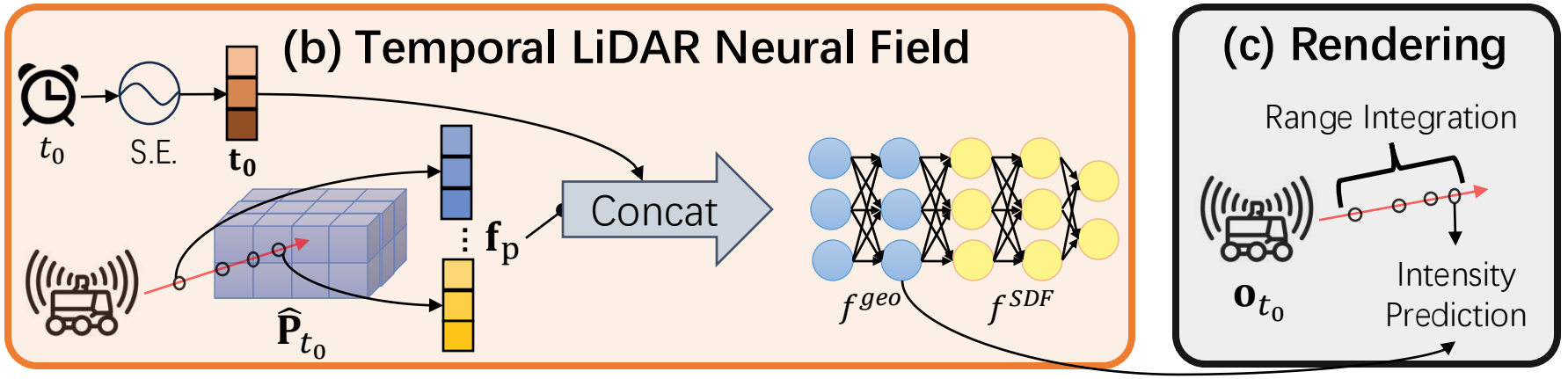
Method: Temporal LiDAR Neural Field

- To represent the 3D scene with embeddings at different timesteps, we propose a Temporal LiDAR Neural Field. For 3D embeddings at different timesteps, we cast LiDAR beams at the ego-vehicle location and sample points along LiDAR beams. Then we interpolate 3D embeddings at the sampled points and concatenate them with time embeddings (sinusoidal encoding). Finally, the concatenated embeddings are fed into f^{geo} and f^{SDF} to predict geometry features and signed distance values.



Method: Point Cloud Rendering

- By aggregating SDF values, we predict range along each LiDAR beam. And we use geometry feature for LiDAR intensity prediction. The final loss function for timestep t_n is L1 loss for predicted range and intensity and also a regularization term on the SDF value on the observed LiDAR point.



$$\mathcal{L}_{t_n} = \frac{1}{N_{\text{render}}} \sum_{i=1}^{N_{\text{render}}} (|r^i - \tilde{r}^i| + |\mathcal{I}^i - \tilde{\mathcal{I}}^i| + |s_i|).$$

Experiment Settings

- To make sure the improvement brought by pre-train is not about accelerating convergence, we first gradually increase the training iterations for randomly initialized models until convergence is observed and then fix the training iterations and use the same schedule for downstream fine-tuning with different pre-training methods.

Experiments

- Pre-trained on Once and fine-tune on Once.

Init.	F.T.	mAP	Vehicle			Pedestrian			Cyclist		
			0-30m	30-50m	50m-	0-30m	30-50m	50m-	0-30m	30-50m	50m-
Ran.	5%	46.07	76.71	51.15	31.84	37.53	20.12	9.84	62.00	42.61	24.18
[73]		44.69 $^{-1.38}$	74.04	49.66	29.63	33.98	20.94	12.42	60.63	43.14	23.63
[74]		44.43 $^{-1.64}$	76.52	49.48	30.18	35.32	18.96	9.36	60.47	40.94	22.99
[22]		40.84 $^{-5.23}$	74.23	46.64	29.45	29.85	17.31	9.56	57.47	33.59	18.34
[35]		45.12 $^{-0.95}$	74.20	49.52	30.25	37.51	20.46	9.97	60.93	41.82	25.75
[21]		46.23 $^{+0.16}$	78.76	55.77	37.81	31.65	16.09	8.78	64.90	44.18	24.73
Ours		**47.84** $^{+1.77}$	79.14	55.68	36.34	35.23	18.00	11.18	64.99	45.80	28.15
Ran.	20%	57.68	82.70	63.37	46.34	52.61	36.48	19.03	71.03	55.34	36.34
[73]		56.27 $^{-1.41}$	81.01	61.13	43.63	49.78	35.51	20.02	69.55	52.58	34.94
[74]		57.09 $^{-0.59}$	83.51	62.57	46.28	50.96	34.55	17.90	70.37	54.50	36.79
[22]		54.30 $^{-3.38}$	80.69	58.95	42.13	45.09	33.14	18.04	68.90	52.20	35.09
[35]		57.23 $^{-0.45}$	81.66	62.64	45.14	51.32	34.80	17.26	70.87	54.08	33.25
[21]		58.08 $^{+0.40}$	84.23	65.44	48.65	49.48	34.84	19.38	70.76	55.75	38.89
Ours		**58.93** $^{+1.25}$	84.08	65.80	50.51	50.31	33.37	19.42	72.54	56.31	39.26
Ran.	100%	65.03	88.18	74.23	61.75	57.32	38.90	21.96	78.07	64.32	48.16
[73]		64.19 $^{-0.84}$	86.07	72.44	59.28	57.25	37.14	22.25	77.62	61.94	45.91
[74]		65.10 $^{+0.07}$	88.02	74.01	61.95	57.56	38.43	22.45	79.95	63.64	47.89
[22]		64.48 $^{-0.55}$	88.34	74.20	61.32	55.78	37.14	22.32	77.95	62.42	46.40
[35]		65.25 $^{+0.22}$	88.31	72.67	62.87	57.48	39.55	24.30	77.92	63.07	48.34
[21]		65.19 $^{+0.16}$	88.11	74.00	62.28	57.67	38.49	21.99	79.51	64.40	47.65
Ours		**66.09** $^{+1.06}$	88.56	75.02	63.10	57.83	39.29	20.63	79.48	65.08	49.02

Experiments

- Transferring pre-trained models to new dataset (Waymo).

Init.	Level-1		Level-2		$\bar{\Delta}$
	mAP	mAPH	mAP	mAPH	
Ran.	61.60	58.58	55.62	52.87	
[21]	61.57	58.57	55.60	52.83	**-0.03**
Ours	62.32	59.22	56.37	53.84	**+0.77**

Experiments

- **Ablation Study** shows the effectiveness of each component

Rec. Emb.	N. F.	Tem. L. N. F.	mAP	NDS
✗	✗	✗	31.06	44.75
✗	✓	✗	32.16	45.26
✓	✓	✗	32.45	45.76
✓	✗	✓	33.17	46.21

Experiments

- **Visualization with moving object labels.**

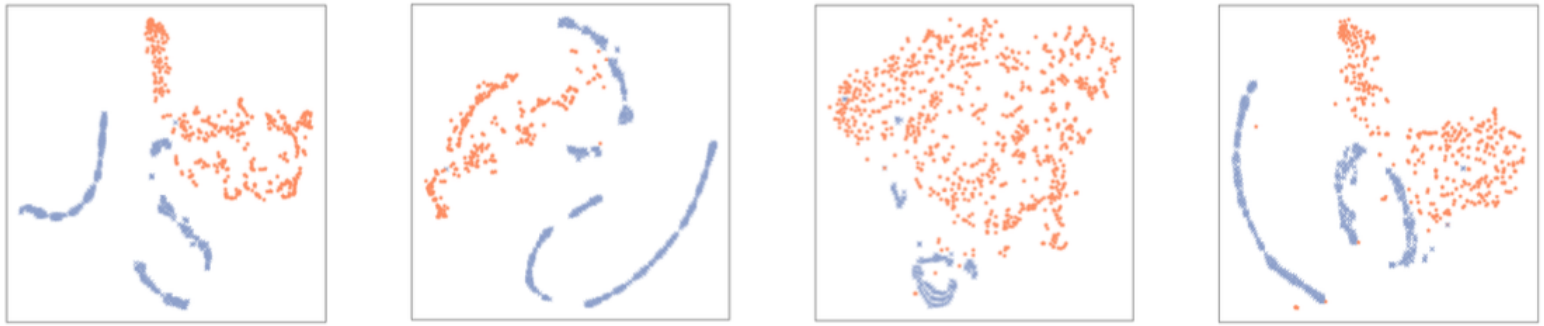


Figure 3: T-SNE visualization of TREND's features with Moving Object labels. The orange ones are static points while grey blue ones are moving points.

Welcome to connect!

Paper

Code

Personal Website