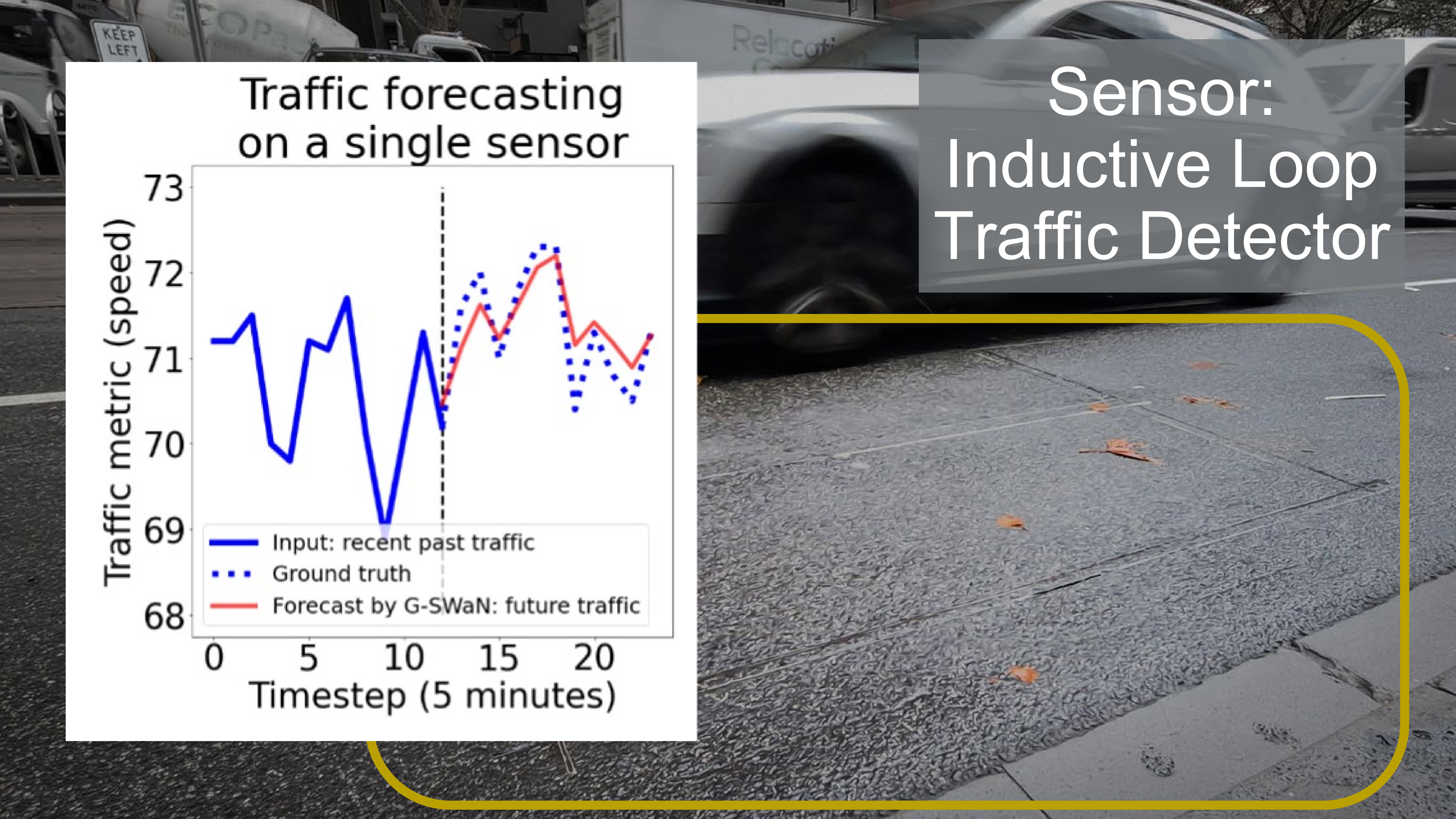
Traffic Forecasting on New Roads Unseen in the Training Data Using Spatial Contrastive Pre-Training.

Arian Prabowo, Hao Xue, Wei Shao, Piotr Koniusz, and Flora D. Salim.

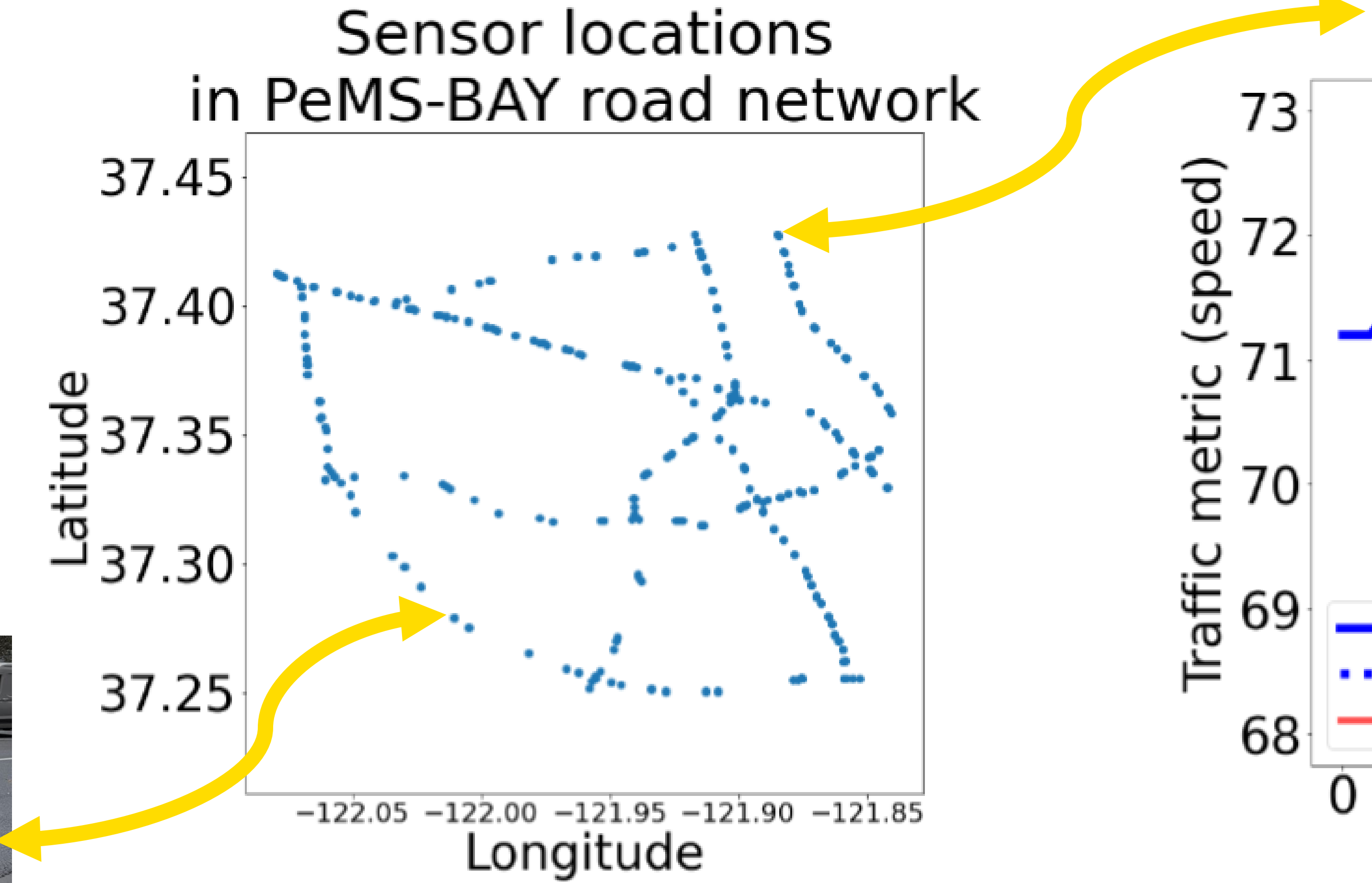
UNSW
SYDNEY

Traffic Forecasting on New Roads Unseen in the Training Data Using Spatial Contrastive Pre-Training.

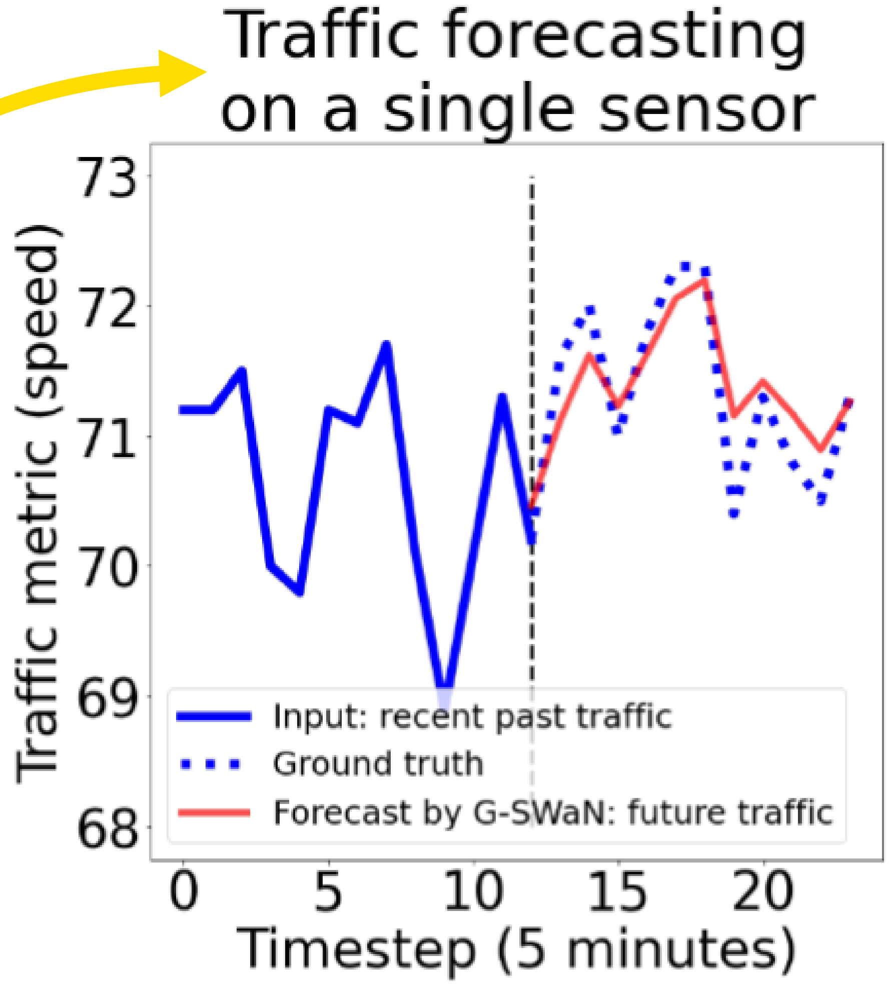
Sensor: Inductive Loop Traffic Detector



KEEP LEFT
Traffic forecasting on a single sensor
Traffic metric (speed)
73
72
71
70
69
68
0
5
10
15
20
Timestep (5 minutes)
Input: recent past traffic
Ground truth
Forecast by G-SWaN: future traffic
Sensor: Inductive Loop Traffic Detector



(a) Locations of the sensors on the Californian highway network surrounding the bay area. Installing a network of sensors on a road infrastructure enables traffic forecasting and smarter cities.



(b) At each sensor, traffic forecasting uses the recent sensor readings (solid blue line) to predict the future traffic (red line). This forecast is made by our proposed architecture Graph Self-attenion WaveNet (G-SWaN). Our forecasts accurately predict the future traffic (dotted blue line).

Fig. 1. Visual abstract of the traffic forecasting task.

What if there is a new road unseen during training?

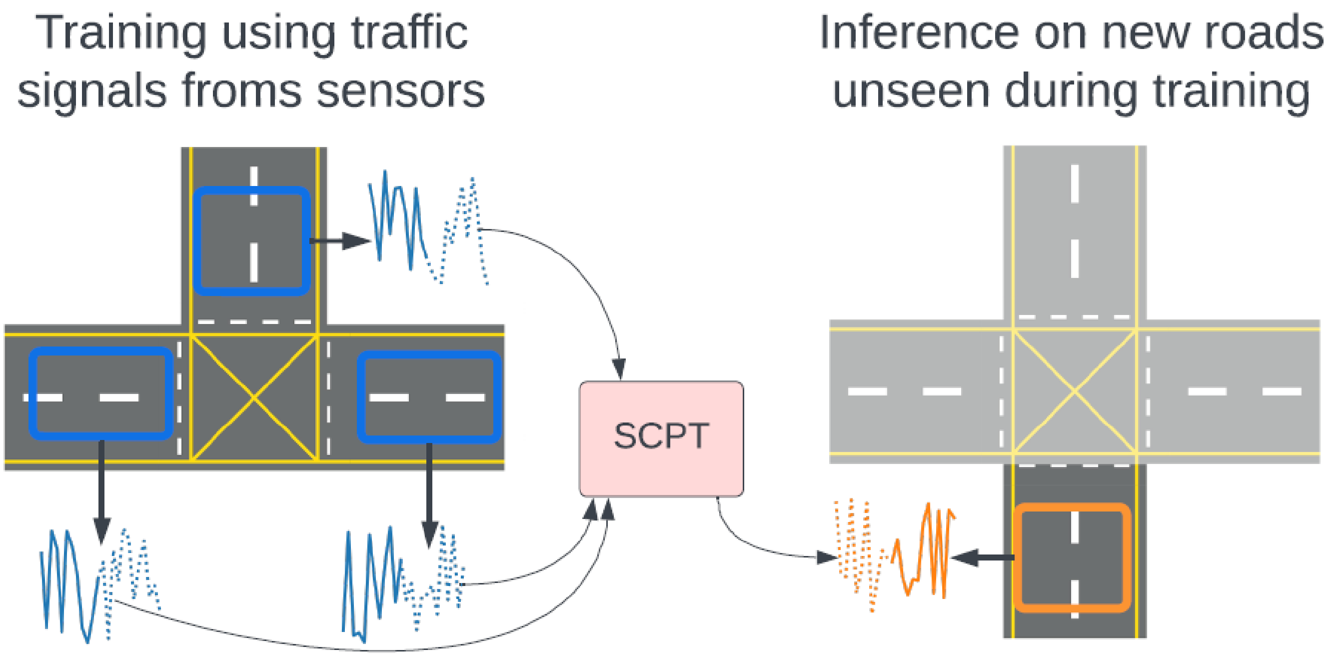


Fig. 1: Our novel traffic forecasting framework, Spatial Contrastive Pre-Training (SCPT), enables accurate forecasts on new roads (orange) that were not seen during training.

<u>Challenge</u>:

A new road unseen during training?

<u>Solution</u>:

A new paradigm:

Spatial Contrastive Pre-Training

(SCPT)

Related works

Related works

- Traffic forecasting are getting very popular, the number of papers grow every year.
- However, the topic is getting saturated.
  - The improvements is very small, very close to be considered as solved; of no interest to actual traffic planners, managers, and engineers.
  - My favorite paper title: Eric L. Manibardo, Ibai Laña, and Javier Del Ser. 2022. **Deep Learning for Road Traffic Forecasting: Does it Make a Difference?** Trans. Intell. Transport. Sys. 23, 7 (July 2022), 6164–6188. https://doi.org/10.1109/TITS.2021.3083957
  - This is turning into a mere intellectual exercise.

Related works

- Traffic forecasting are getting very popular, the number of papers grow every year.
- However, the topic is getting saturated.
- The datasets are artificially small

Table 4: Detailed statistics on the real world datasets.

	Dataset:	METR-LA	PeMS-BAY	PeMS-D7(m)	PeMS-11k(s)
Spatial	Nodes	207	325	228	11,160
	Edges	1,515	2,694	7,304	234,966

Related works

- Traffic forecasting are getting very popular, the number of papers grow every year.
- However, the topic is getting saturated.
- The datasets are artificially small
  - Only 1 prior work (to our knowledge) tried to tackle the 11k dataset. They used graph partitioning. Mallick, T., Balaprakash, P., Rask, E., Macfarlane, J.: Graph-partitioning based diffusion convolutional recurrent neural network for large-scale traffic forecasting. Transportation Research Record 2674(9), 473–488 (2020)

Related works

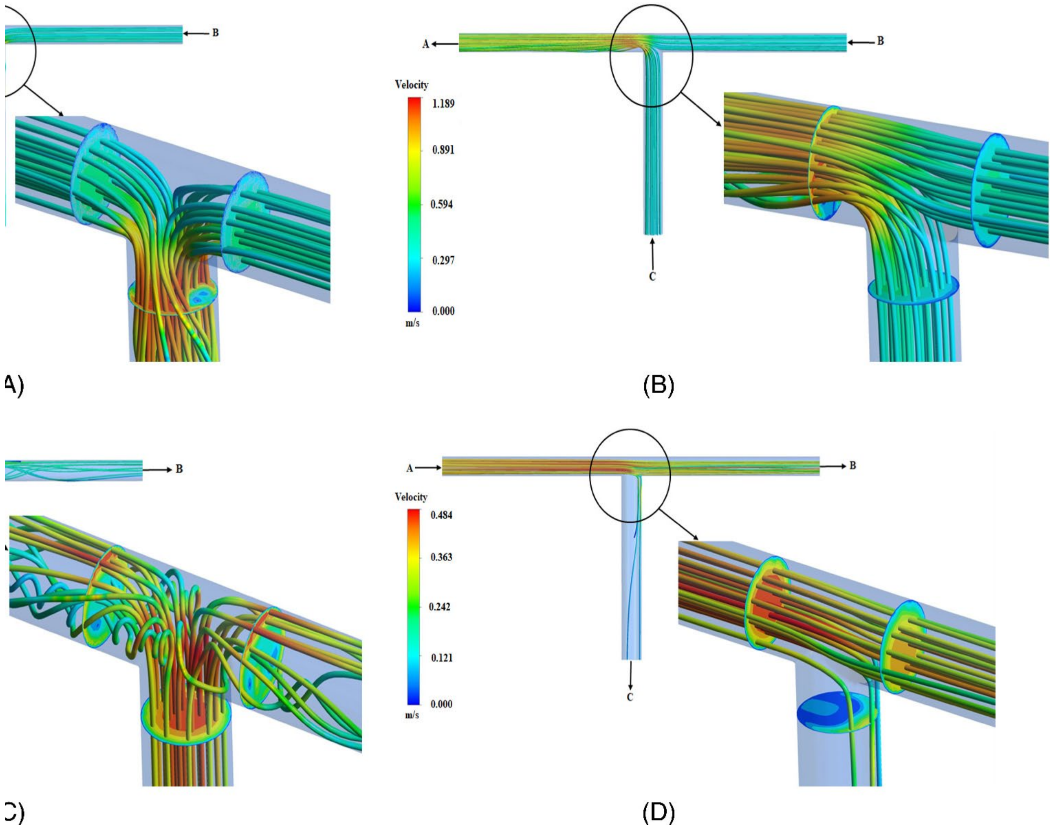
- Traffic forecasting are getting very popular, the number of papers grow every year.
- However, the topic is getting saturated.
- The datasets are artificially small
  - Only 1 prior work (to our knowledge) tried to tackle the 11k dataset. They used graph partitioning. Mallick, T., Balaprakash, P., Rask, E., Macfarlane, J.: Graph-partitioning based diffusion convolutional recurrent neural network for large-scale traffic forecasting. Transportation Research Record 2674(9), 473–488 (2020)
  - Scaling is not trivial

Scaling is not trivial

- Many prior works (and many works presented in this week in EMCL PKDD 2023) find **<u>node embedding</u>** to be very effective.

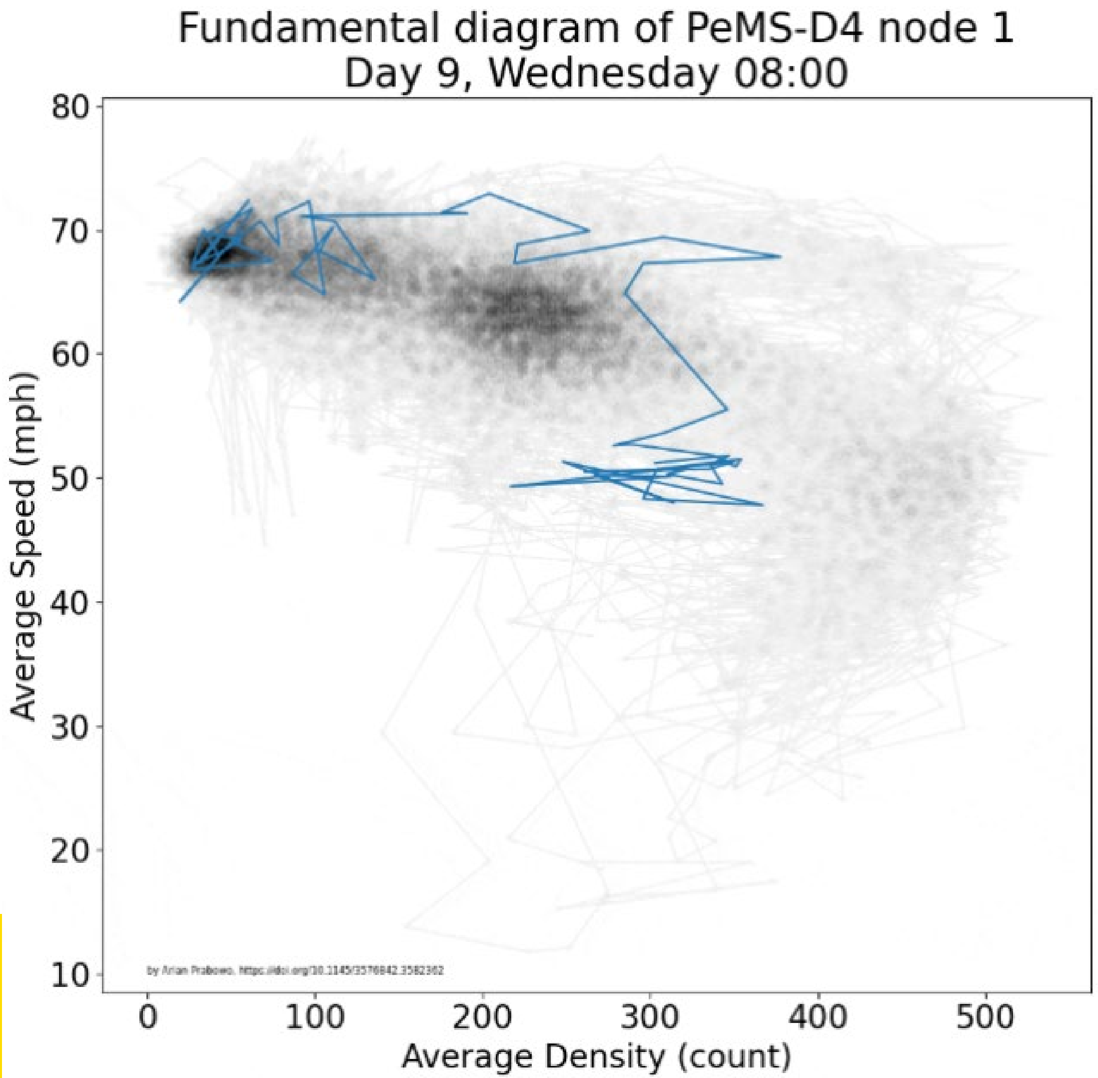
But why is node embedding so effective?

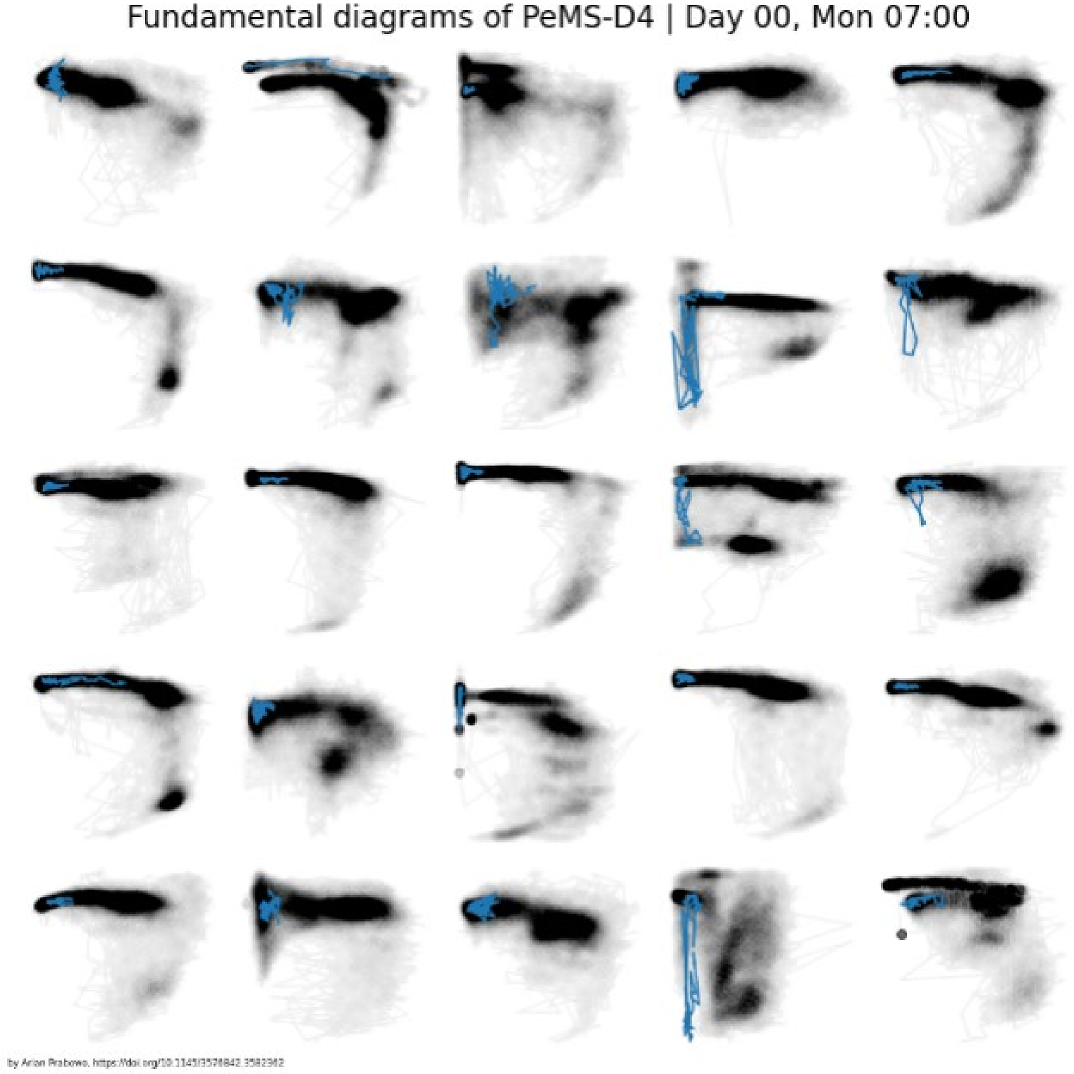
- One classical way to model traffic is as fluid, using PDE (this ties back to today's keynote: PINN)
- The PDE, explicitly as equations or implicitly as **<u>the neural network, is symmetric</u>** across the entire traffic network.
- From this perspective, **<u>the node embeddings act as a boundary conditions</u>**



Gajbhiye, BD, Kulkarni, HA, Tiwari, SS, Mathpati, CS. Teaching turbulent flow through pipe fittings using computational fluid dynamics approach. *Engineering Reports*. 2020; 2:e12093. https://doi.org/10.1002/eng2.12093

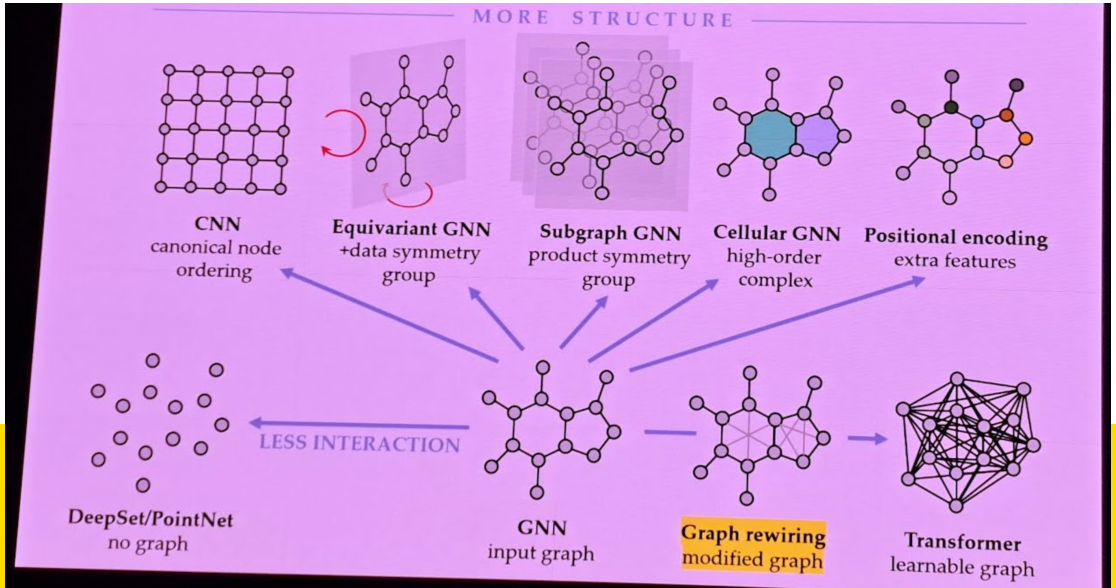
Our prior work showing that **<u>every node is unique</u>**, emphasizing the need for node embedding.





Scaling is not trivial

- Many prior works (and many works presented in this week in EMCL PKDD 2023) find **node embedding** to be very effective.
- Many prior works, and as explained during the keynote this morning, **graph re-wiring** is important.



Scaling is not trivial

- Many prior works (and many works presented in this week in EMCL PKDD 2023) find **node embedding** to be very effective.
- Many prior works, and as explained during the keynote this morning, **graph re-wiring** is important. However, most of the current graph re-wiring implemented in traffic forecasting (usually called as adaptive adjacency matrix) has the complexity of $O(n^2)$.

Related works

- Traffic forecasting are getting very popular, the number of papers grow every year.
- However, the topic is getting saturated.
- The datasets are artificially small
- **<u>Contrastive learning</u>** is struggling to find popularity in traffic forecasting.

Related works

- Traffic forecasting are getting very popular, the number of papers grow every year.
- However, the topic is getting saturated.
- The datasets are artificially small
- Contrastive learning is struggling to find popularity in traffic forecasting.
- Only one prior work (**FUNS-N**) that tried to solve unseen roads, however they are context-driven instead of data driven.
  - Roth, A., Liebig, T.: Forecasting unobserved node states with spatiotemporal graph neural networks. Data Mining Workshops ICDMW'22 (2022)

Research Gap

- Classical traffic forecasting are close to be considered as “solved”.
- Difficulty in implementing contrastive learning in traffic.
- Small dataset.

Our Contributions

- We propose a new task: forecasting on unseen roads.
- We successfully implement contrastive learning to address the new task.
- We can use this paradigm to scale forecasting to a very large network (10k nodes)

Methods

UNSW
SYDNEY

Methods (quick glance)

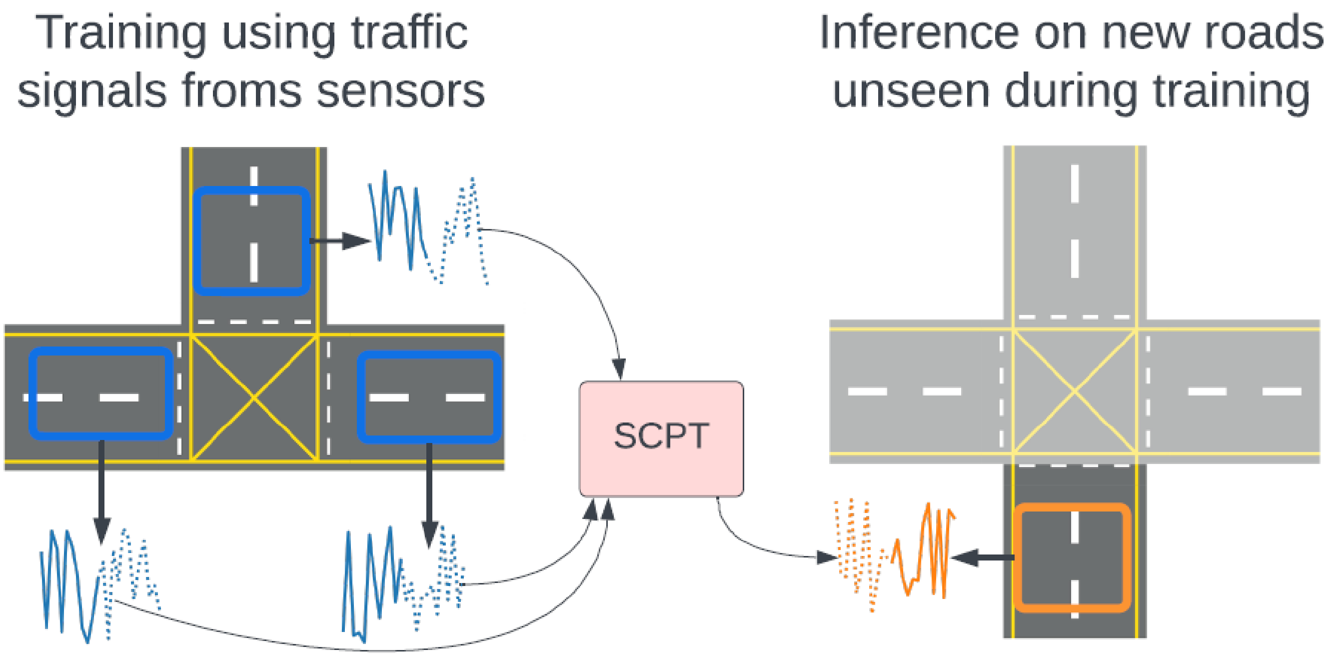


Fig. 1: Our novel traffic forecasting framework, Spatial Contrastive Pre-Training (SCPT), enables accurate forecasts on new roads (orange) that were not seen during training.

We pre-trained a spatial encoder using SCPT.

During inference time, it infers the spatial embedding of new roads from minimal data.

3.2 Spatio-temporal split

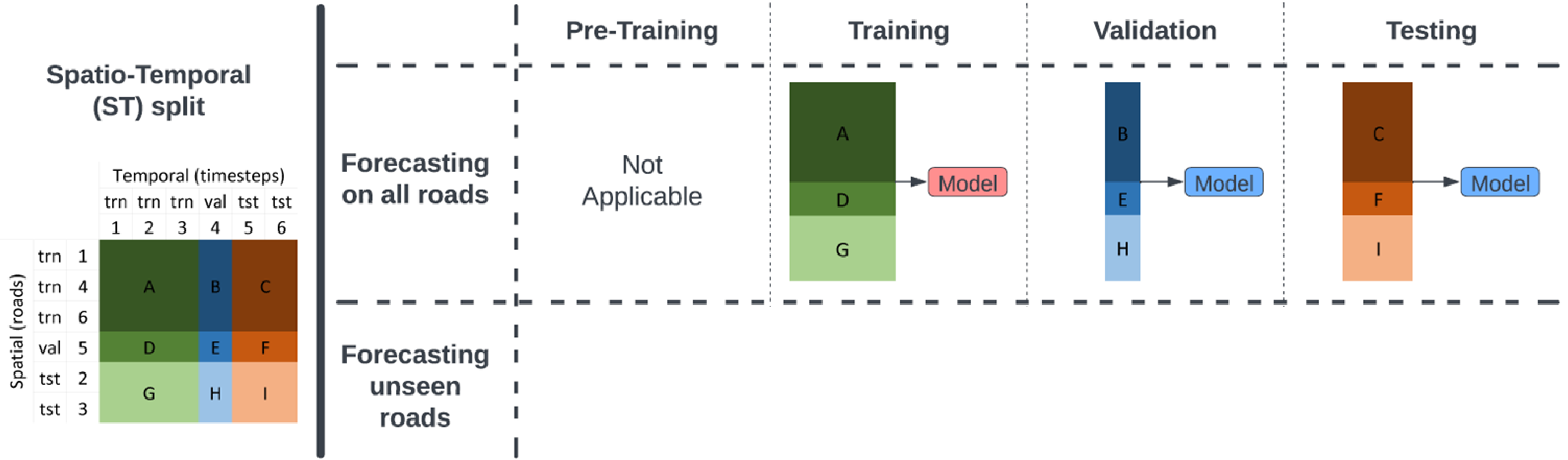


Fig. 2: The ST splitting strategy divides the dataset into nine subsets (left side), while the right side illustrates the usage of different subset combinations at different stages.

3.2 Spatio-temporal split

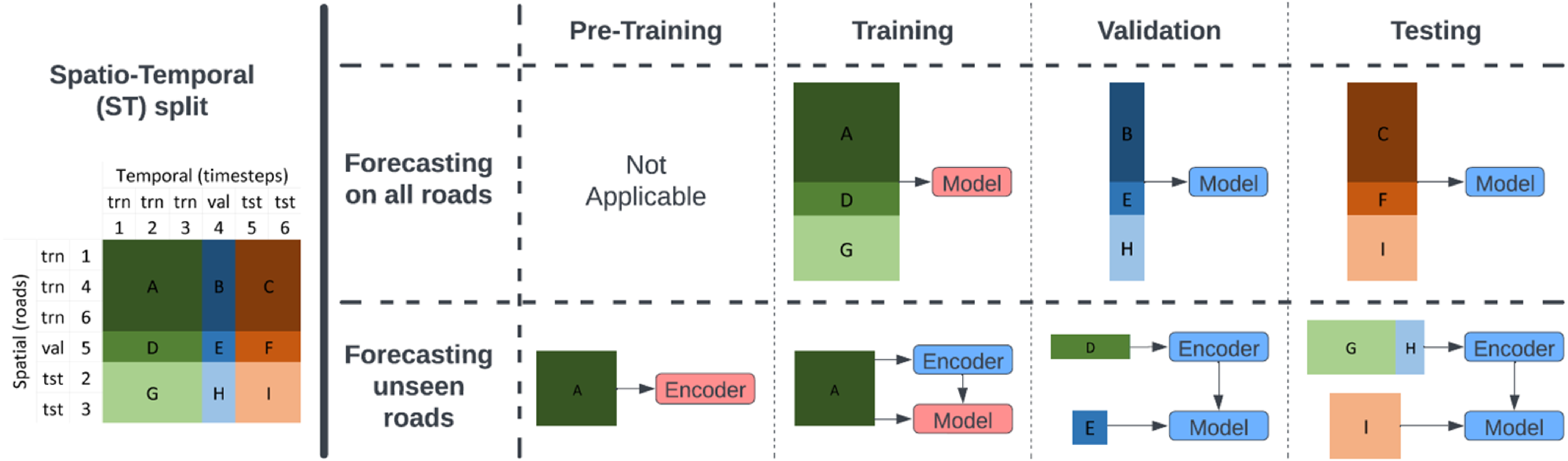


Fig. 2: The ST splitting strategy divides the dataset into nine subsets (left side), while the right side illustrates the usage of different subset combinations at different stages.

This is SimCLR like, very popular in CV.

However, the encoder is stochastic

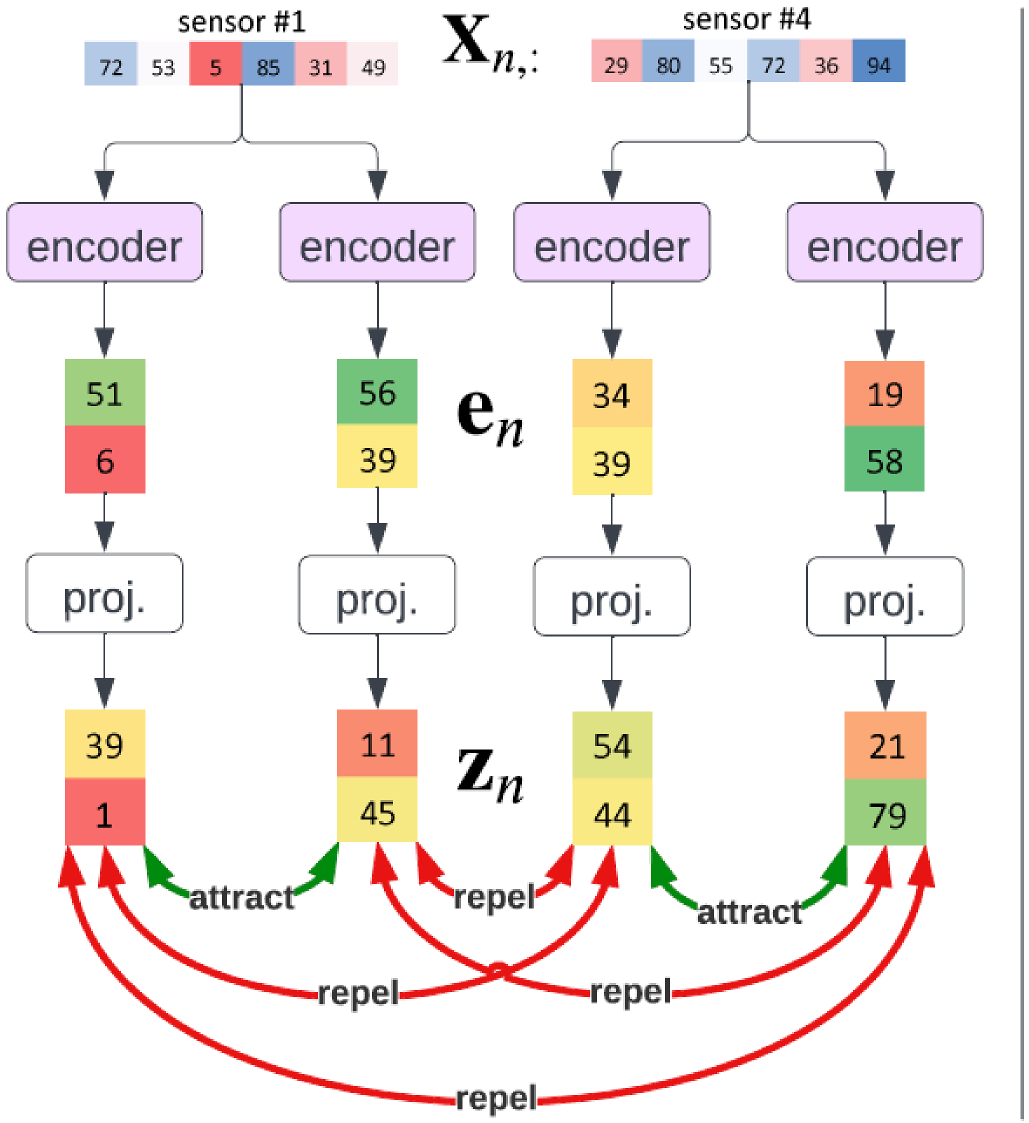
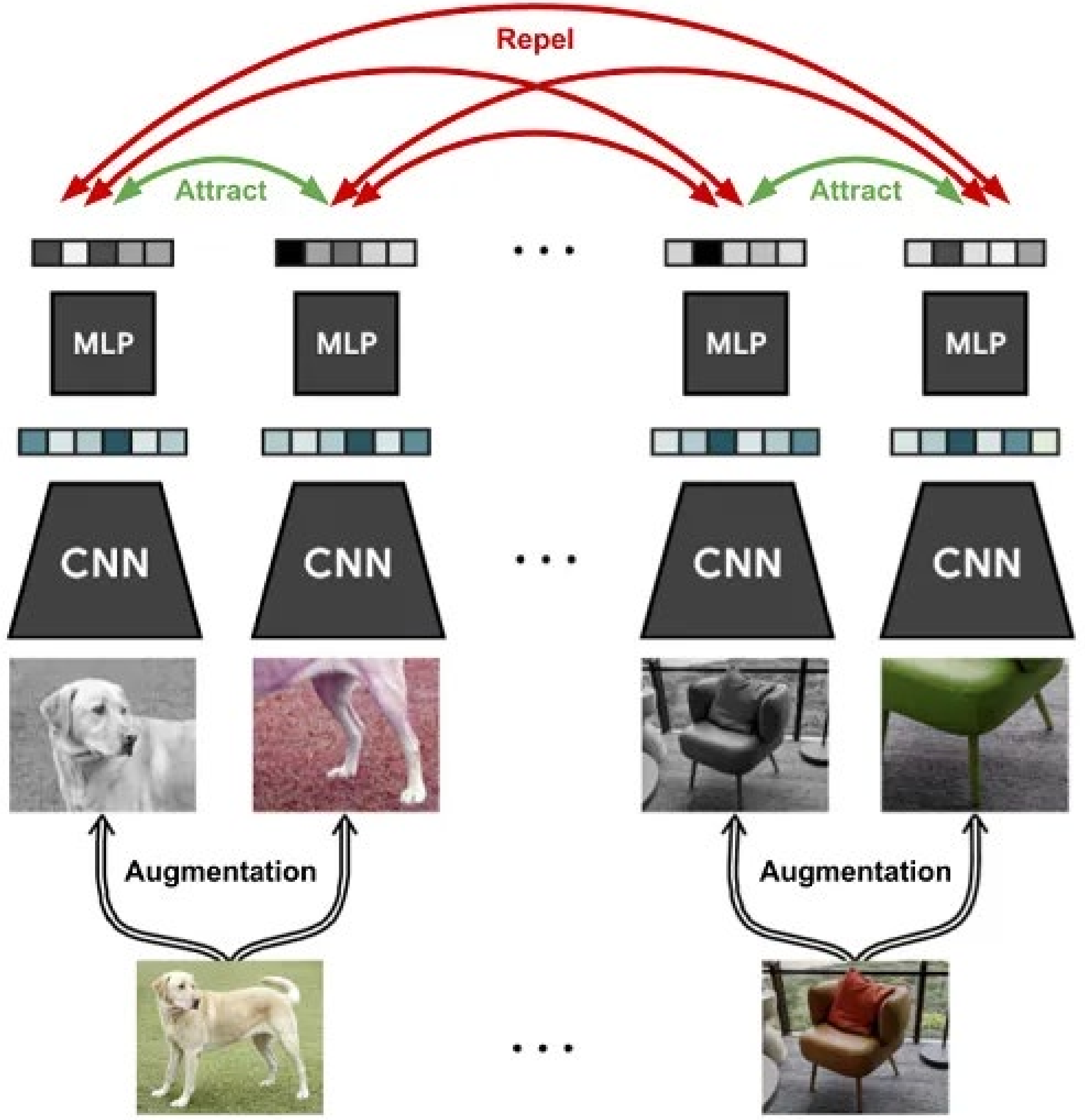
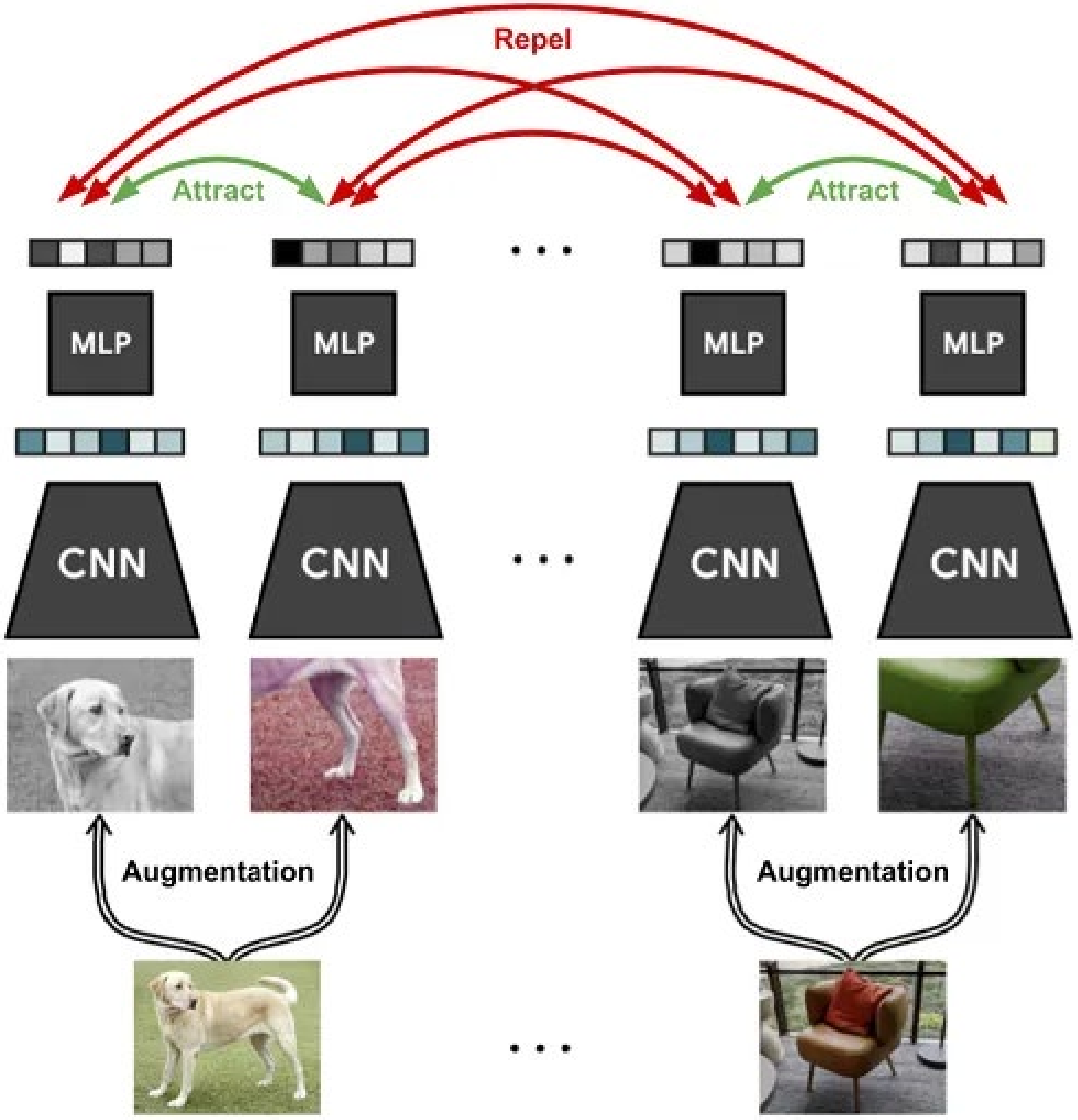


Fig. 3: On the left, the use of contrastive loss to pre-train the spatial encoder is depicted, while on the right, the framework of the (spatial) encoder is illustrated.

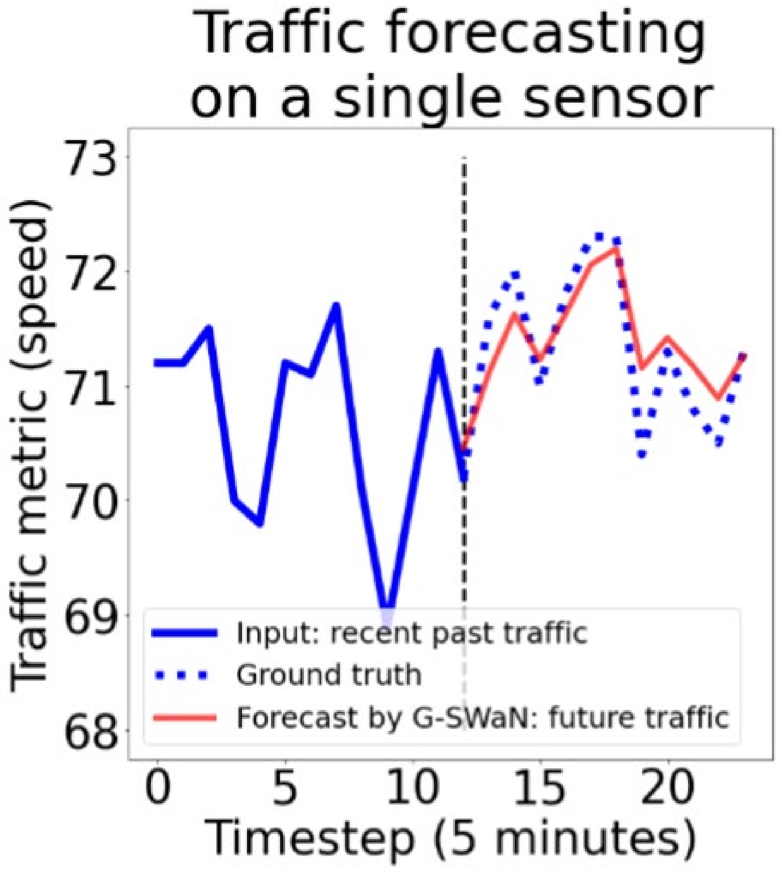


Original Sim-CLR

https://github.com/google-research/simclr



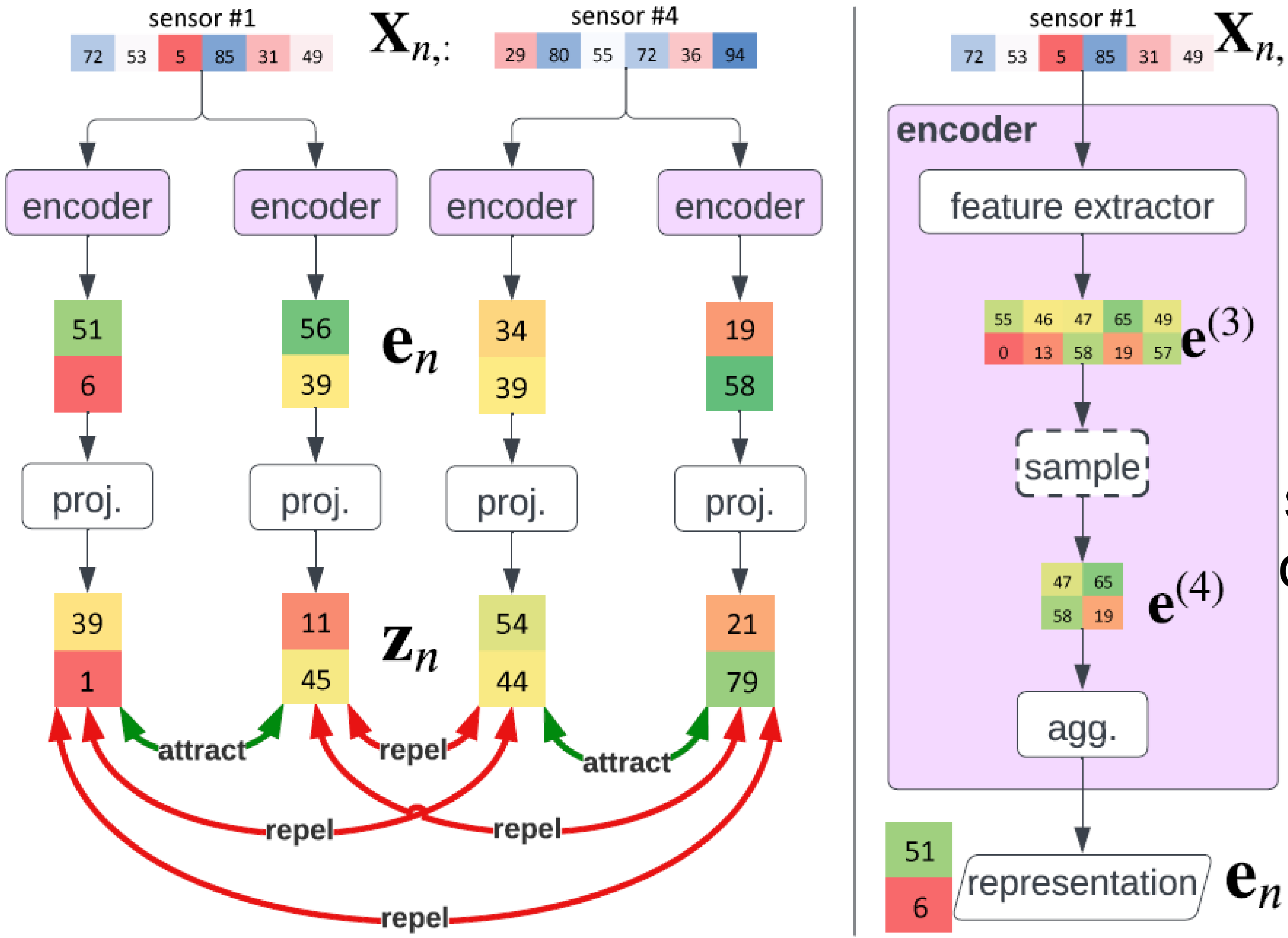
Increasing brightness (speed) by 10 mph means everyone is breaking the law.



The encoder is stochastic.

We replace augmentation with sampling.

Sampling also allows input of variable length.



Because of this sampling operation

Fig. 3: On the left, the use of contrastive loss to pre-train the spatial encoder is depicted, while on the right, the framework of the (spatial) encoder is illustrated.

We used Graph WaveNet as the backbone

But this is architecture agnostic

Only that such architecture uses node embeddings

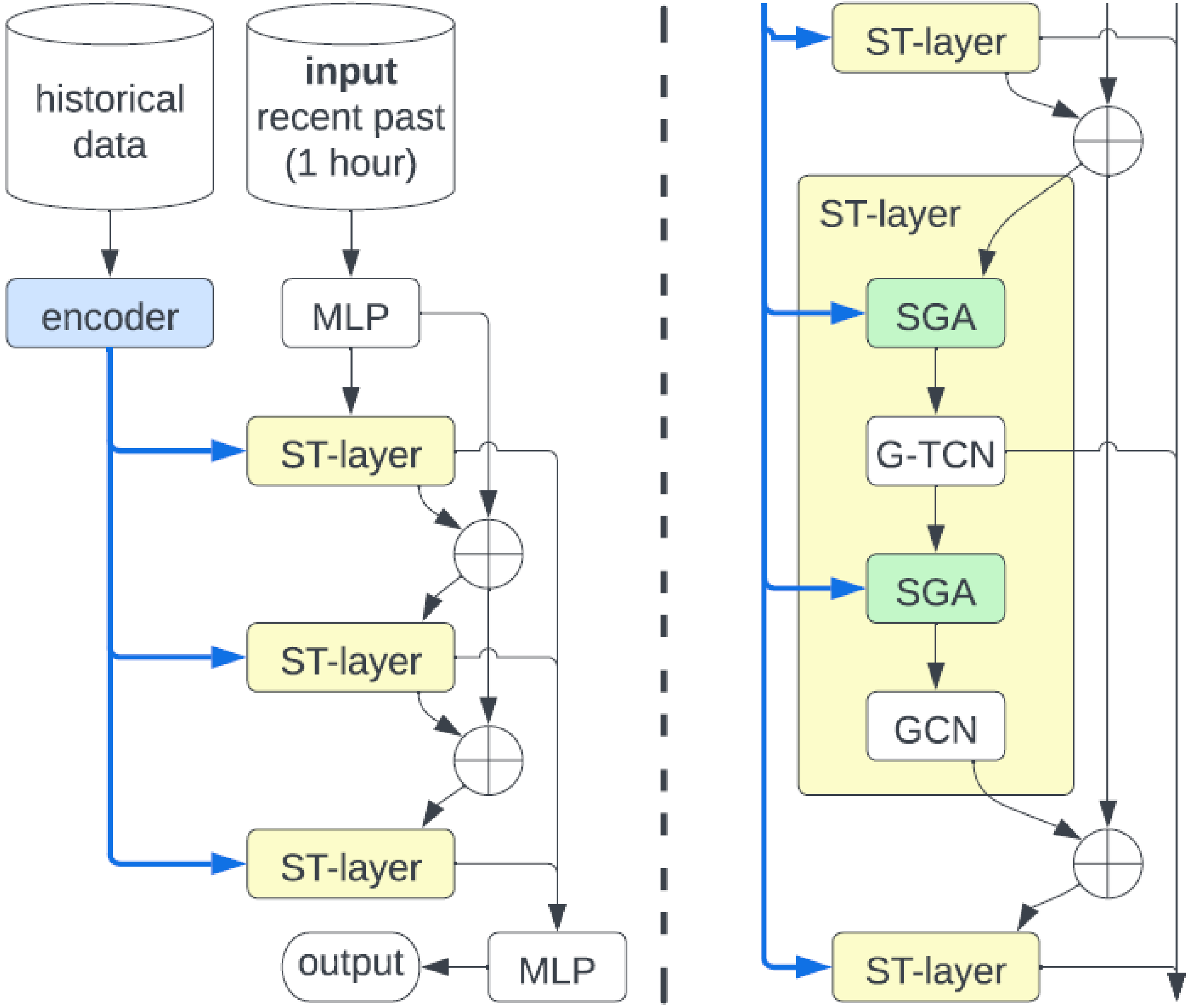


Fig. 4: The left side of the figure illustrates the flow of outputs from the spatial encoder (blue) into the spatio-temporal (ST) layers (yellow). On the right side, the usage of Spatially Gated Addition (SGA) to integrate spatial information from the spatial encoder into the input of the G-TCN and GCN layers within the ST-layer is depicted.

Results

Table 1: Performances evaluation of the SCPT framework using ST split. In this setup, the models are trained on only 70%, validated on 10%, and tested on 20% of the roads. This table shows the average performance across 12 timesteps (1 hour) on the 20% of the roads that are unseen during the training. $\Delta(\%)$ denotes the percentage of error reduction.

Dataset	Methods	RMSE	MAE	MAPE
METR-LA	GWN	10.3405 ± 0.2634	4.7373 ± 0.1618	12.2677 ± 0.8058
	GWN+SCPT	**10.0385** ± 0.2112	**4.5645** ± 0.1556	**11.5002** ± 0.8007
	$\Delta(\%)$	3%	4%	6%
PeMS-BAY	GWN	4.5059 ± 0.1613	2.0126 ± 0.1037	4.7779 ± 0.4303
	GWN+SCPT	**3.9658** ± 0.1266	**1.8163** ± 0.0875	**4.1358** ± 0.2740
	$\Delta(\%)$	12%	10%	13%
PeMS-D7(m)	GWN	6.4635 ± 0.3103	3.4327 ± 0.1974	8.6896 ± 0.7844
	GWN+SCPT	**5.6893** ± 0.2552	**3.0794** ± 0.1448	**7.6770** ± 0.6678
	$\Delta(\%)$	12%	10%	12%

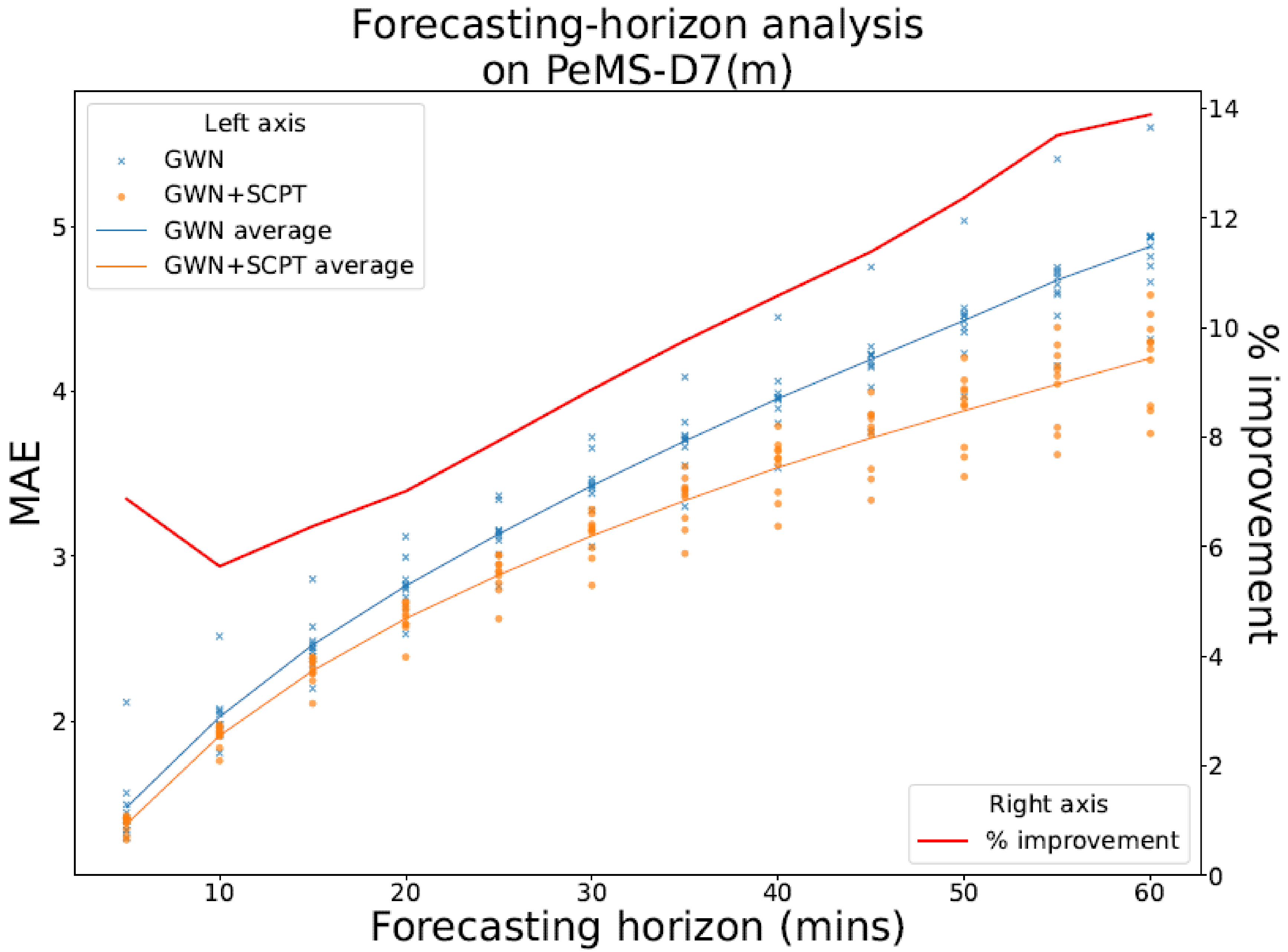


Fig. 5: Performance across forecasting horizons.

Forecasting on large real-world road network

Table 4: Detailed statistics on the real world datasets.

	Dataset:	METR-LA	PeMS-BAY	PeMS-D7(m)	PeMS-11k(s)
Spatial	Nodes	207	325	228	11,160
	Edges	1,515	2,694	7,304	234,966
Temporal	Duration (timesteps)	34,272	52,116	12,672	25,632
	Duration (days)	121	150	61	89
	Time start	01-Mar-12	01-Jan-17	01-May-12	01-Feb-18
	Time end	30-Jun-12	31-May-17	30-Jun-12	30-Apr-18
	Granularity (mins)	5	5	5	5
Speed (mph)	Min	0.00	0.00	3.00	3.00
	Q1	57.13	62.10	57.50	62.60
	Median	63.22	65.30	64.10	65.10
	Mean	58.46	62.62	58.89	63.14
	Q3	66.50	67.50	66.70	67.80
	Max	70.00	85.10	82.60	99.30
	Standard Deviation	20.26	9.59	13.48	9.01
	Missing values	8.82%	0.00%	0.00%	0.00%
Size	Entry	7,094,304	16,937,700	2,889,216	286,053,120
	Compressed (MB)	54	130	6	2,235

Most traffic dataset is artificially small.

A sub network is selected for research only.

Forecasting on large real-world road network

1. Use one large model
   a) Inefficient
   b) More technical overhead e.g.: distributed data parallel
   c) Scaling issues:
      a) How find the node embedding of unseen roads?
      b) $O(n^2)$ cost for graph re-wiring / adaptive adjacency

Forecasting on large real-world road network

1. Use one large model
2. Use graph partitioning, and train a model for each partition [GP-DCRNN] .

[GP-DCRNN] Mallick, T., Balaprakash, P., Rask, E. and Macfarlane, J., 2020. Graph-partitioning-based diffusion convolutional recurrent neural network for large-scale traffic forecasting. *Transportation Research Record*, *2674*(9), pp.473-488.

Forecasting on large real-world road network

1. Use one large model
2. Use graph partitioning, and train a model for each partition [GP-DCRNN] .
3. Train only on a small (1%) of the roads, and treat the 99% as new roads.

Table 3: Performance comparison on using the SCPT framework to train on a small sample (1%) of roads to scale to a large dataset PeMS-11k(s). Δ(%) denotes the percentage of error reduction.

Method:	GWN	GWN+SCPT	Δ(%)	GP-DCRNN
RMSE	5.6345 ± 0.7469	**4.6741** ± 0.2089	17%	
MAE	2.8241 ± 0.2840	**2.4273** ± 0.2171	14%	
MAPE	5.6345 ± 0.7469	**4.6741** ± 0.2089	17%	
medianMAE12	3.4554 ± 0.2343	3.2442 ± 0.3071	6%	**2.0200**
Training time	**00:16:39**	00:22:28		7 days, 22:34:53
Roads seen in training (count)	111			11160
Roads seen in training (%)	1%			100%

Favorable trade-off between error and speed.

UNSW SYDNEY

Conclusion

- Traffic forecasting on **unseen roads** is an interesting new task.
  - From theoretical Graph-ML and geometric DL perspective.
  - From PINN perspective (traffic is a PDE flow).
  - From mobility / spatiotemporal / timeseries perspective.
  - From applied data science: traffic planners and engineers.
- **SCPT** is the first data-driven solution to this task.
- **Scales efficiently** to large (11k nodes) network.
  - Also allows engineers to adjust the **trade-off** between resource use and performance.

UNSW SYDNEY

Thank You. Any Questions?

Traffic Forecasting on New Roads Unseen in the Training Data Using Spatial Contrastive Pre-Training (SCPT).

Link to GitHub

https://github.com/cruiseresearchgroup/forecasting-on-new-roads

UNSW SYDNEY

CSIRO

RMIT UNIVERSITY

Arian Prabowo,
Hao Xue,
Wei Shao,
Piotr Koniusz, and
Flora D. Salim.

Table 2: Ablation study on the PeMS-D7(m) dataset. Every experiment is replicated five times (except the first and last ones). The first row is the backbone baseline GWN and the last row is the full GWN+SCPT. The + column shows the MAE reduction when compared to the GWN baseline (first row). The - column shows the performance reduction when compared to the full SCPT framework.

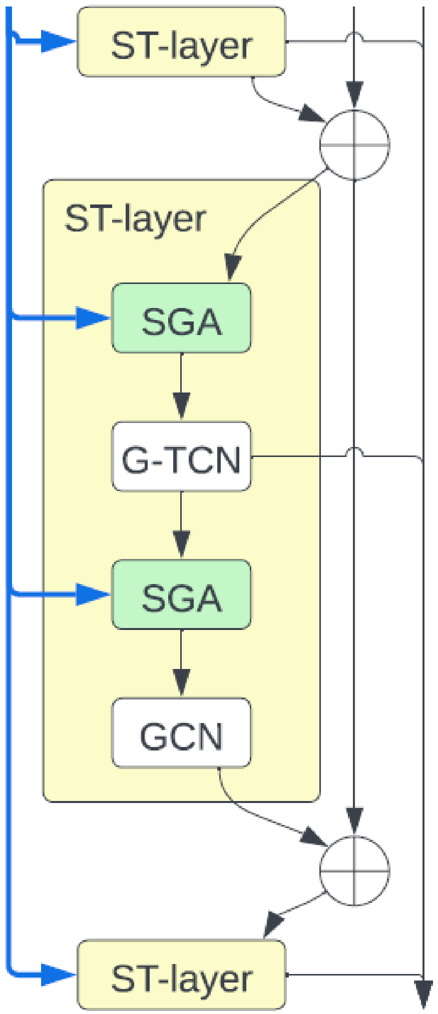
Methods				MAE			
SCPT	SGA	Decoupling	AdpAdj	mean	std.	+	-
0	0	0	0	3.433	0.197	0.000	0.353
✓	0	0	0	3.366	0.181	0.066	0.287
✓	✓	0	0	3.349	0.161	0.083	0.270
✓	0	✓	0	3.398	0.236	0.035	0.319
✓	0	0	✓	3.350	0.234	0.083	0.271
✓	0	✓	✓	3.249	0.255	0.184	0.169
✓	✓	0	✓	3.101	0.141	0.332	0.022
✓	✓	✓	0	3.406	0.187	0.026	0.327
✓	✓	✓	✓	**3.079**	0.145	**0.353**	**0.000**

In brief, $SGA(\cdot)$ layer adds the latent representation vector $\mathbf{e}_n$ to the activation $\mathbf{h}_n^{(l)}$ of the l^{th} layer of the model, weighted by a coefficient $c_n(\mathbf{h}_n^{(l)}, \mathbf{e}_n)$ that is unique for every sensor n. More formally:

$$\mathbf{h}_n^{(l+1)} = SGA(\mathbf{h}_n^{(l)}, \mathbf{e}_n) = \mathbf{h}_n^{(l)} + c_n(\mathbf{h}_n^{(l)}, \mathbf{e}_n)\mathbf{e}_n$$

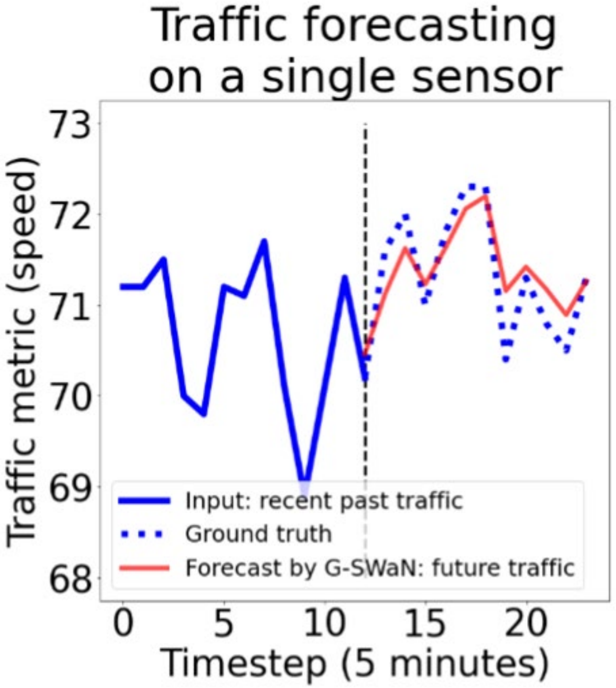
where $\mathbf{h}_n^{(l)} \in \mathbb{R}^D$ is the activation of the l^{th} layer for sensor n in the forecasting model, $\mathbf{e}_n \in \mathbb{R}^D$ is the latent representation of sensor n (the output of the frozen spatial encoder), and $c_n(\cdot) \in \mathbb{R}$ is the weight for sensor n at layer l. We calculate the weight $c_n(\cdot)$ using a multi-layer perceptron (MLP) with one hidden layer, ReLU activation, and wrap it under a sigmoid $\sigma(\cdot)$ to ensure that the weight is between 0 and 1:

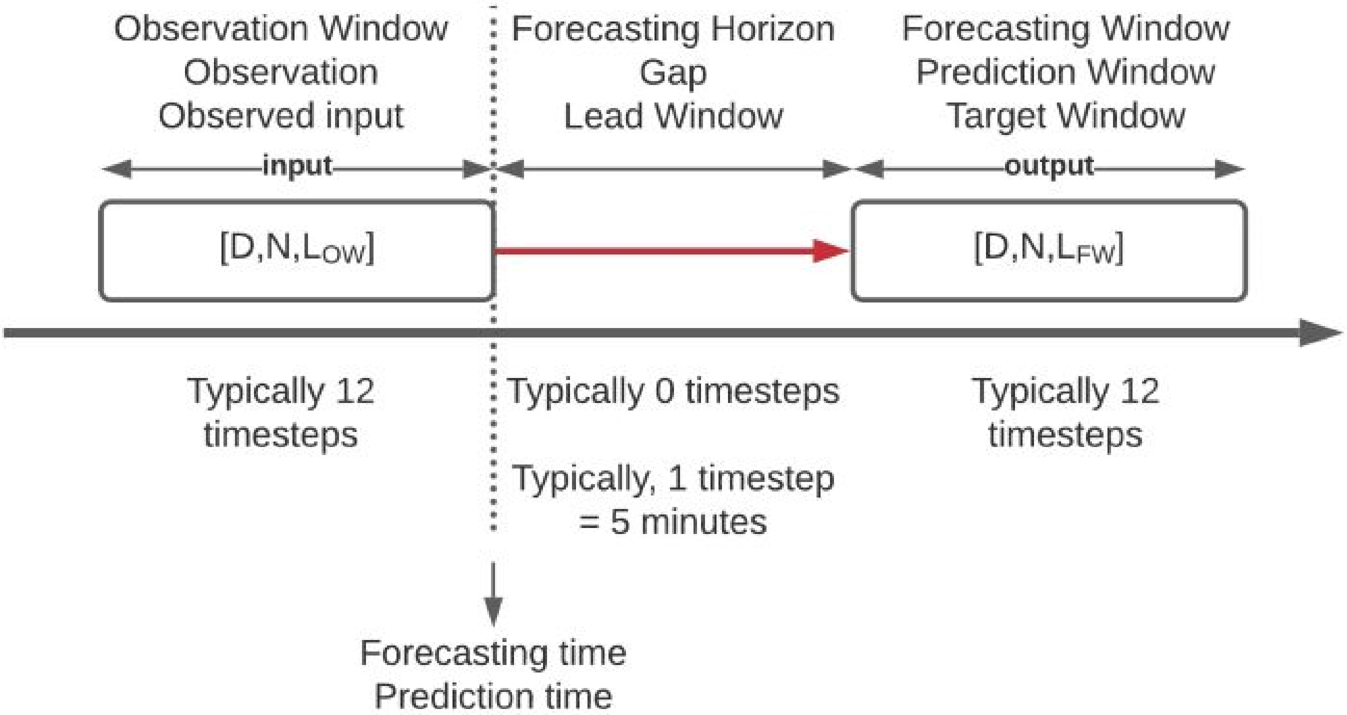
$$c_n(\mathbf{h}_n^{(l)}, \mathbf{e}_n) \in \mathbb{R} = \sigma\left(FC^{(2)}\left(ReLU\left(FC^{(1)}\left(\mathbf{h}_n^{(l)} + \mathbf{e}_n\right)\right)\right)\right).$$



SGA

Fig. 4: The left side of the figure illustrates the flow of outputs from the spatial encoder (blue) into the spatio-temporal (ST) layers (yellow). On the right side, the usage of Spatially Gated Addition (SGA) to integrate spatial information from the spatial encoder into the input of the G-TCN and GCN layers within the ST-layer is depicted.





Traffic Forecasting: Problem Definition

MAE		spatial seed 1	spatial seed 2	spatial seed 3	spatial seed 4	spatial seed 5	spatial seed 6	std. (0.060)
model initialization seed	1	4.713	4.660	4.773	4.698	4.634	4.718	0.044
	2	4.679	4.778	4.727	4.698	4.702	4.756	0.034
	3	4.598	4.684	4.727	4.620	4.788	4.692	0.063
	4	4.815	4.706	4.777	4.654	4.668	4.766	0.059
	5	4.647	4.723	4.724	4.707	4.577	4.731	0.056
	6	4.872	4.608	4.698	4.676	4.535	4.612	0.106
	std. (0.057)	0.095	0.053	0.028	0.031	0.083	0.051	**0.068**

Fig. 6: Analyzing the proposed framework's performance variance based on randomness in sensor selections in comparison with randomness in model weight initialization.

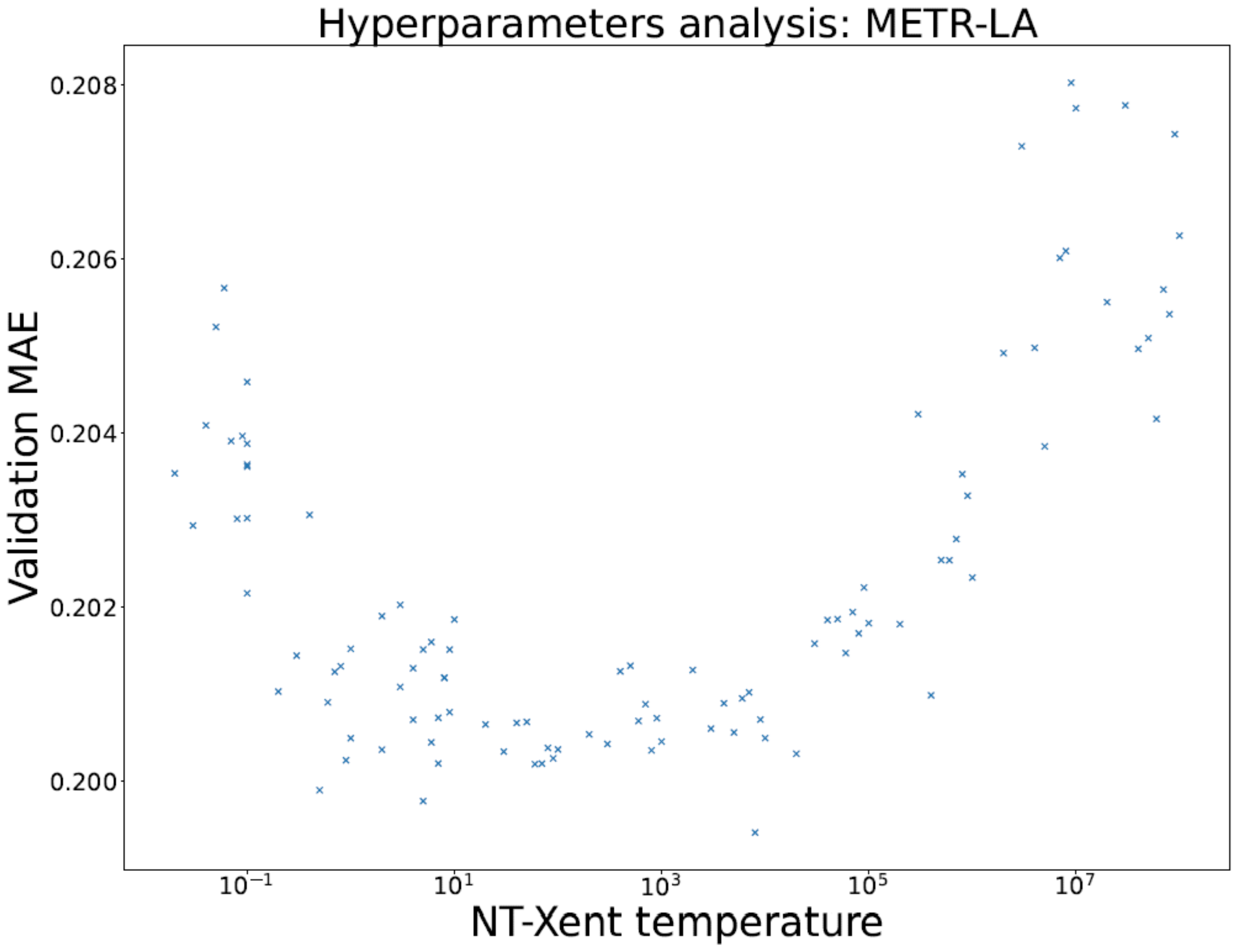


Fig. 7: Analyzing the framework's sensitivity against NT-Xent temperature hyperparameter.

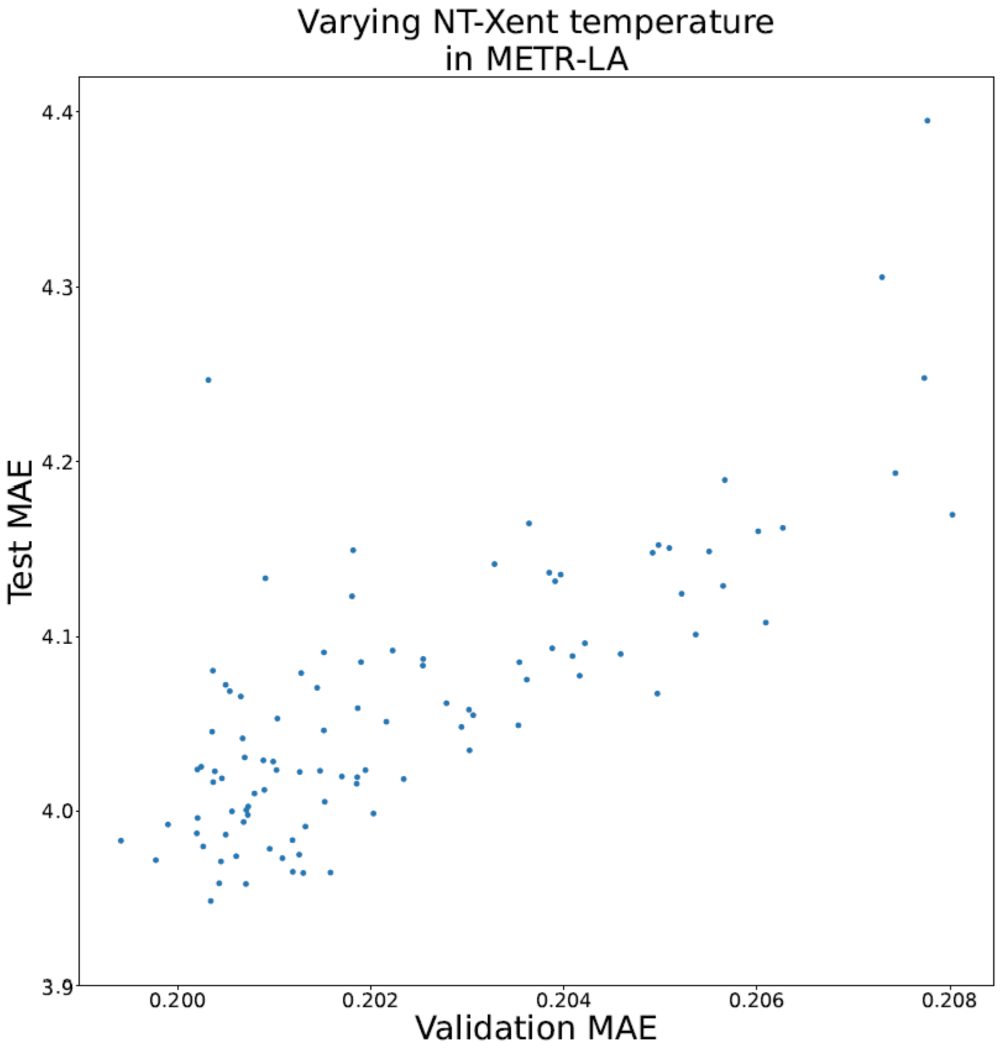


Fig. 8: Correlation validation and test MAE when verying NT-Xent temperature.