Spatio-Temporal Self-Supervised Learning for Traffic Flow Prediction

Jiahao Ji[1], **Jingyuan Wang**[1,2,3*], **Chao Huang**[4],
Junjie Wu[3], **Boren Xu**[1], **Zhenhe Wu**[1], **Junbo Zhang**[5,6], **Yu Zheng**[5,6]

[1]School of Computer Science & Engineering, Beihang University, China
[2]Peng Cheng Laboratory, China [3]School of Economics & Management, Beihang University, China
[4]Department of Computer Science, Musketeers Foundation Institute of Data Science, University of Hong Kong, China
[5]JD Intelligent Cities Research, Beijing, China [6]JD iCity, JD Technology, Beijing, China

Abstract

Robust prediction of citywide traffic flows at different time periods plays a crucial role in intelligent transportation systems. While previous work has made great efforts to model spatio-temporal correlations, existing methods still suffer from two key limitations: i) Most models collectively predict all regions' flows without accounting for spatial heterogeneity, *i.e.,* different regions may have skewed traffic flow distributions. ii) These models fail to capture the temporal heterogeneity induced by time-varying traffic patterns, as they typically model temporal correlations with a shared parameterized space for all time periods. To tackle these challenges, we propose a novel Spatio-Temporal Self-Supervised Learning (ST-SSL[1]) traffic prediction framework which enhances the traffic pattern representations to be reflective of both spatial and temporal heterogeneity, with auxiliary self-supervised learning paradigms. Specifically, our ST-SSL is built over an integrated module with temporal and spatial convolutions for encoding the information across space and time. To achieve the adaptive spatio-temporal self-supervised learning, our ST-SSL first performs the adaptive augmentation over the traffic flow graph data at both attribute- and structure-levels. On top of the augmented traffic graph, two SSL auxiliary tasks are constructed to supplement the main traffic prediction task with spatial and temporal heterogeneity-aware augmentation. Experiments on four benchmark datasets demonstrate that ST-SSL consistently outperforms various state-of-the-art baselines. Since spatio-temporal heterogeneity widely exists in practical datasets, the proposed framework may also cast light on other spatial-temporal applications. Model implementation is available at https://github.com/Echo-Ji/ST-SSL.

1 Introduction

Robust traffic flow prediction across different spatial regions at different time periods is crucial for advancing intelligent transportation systems (Zhang et al. 2020). For example, accurate traffic prediction results can not only enable effective traffic controls in a timely manner, but also mitigate tragedies caused by the sudden traffic flow spike. In general,

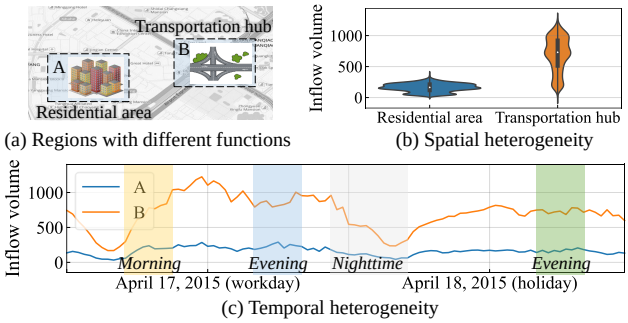


Figure 1: Illustration of our motivation, *i.e.,* the spatial and temporal heterogeneity of traffic flow data.

traffic prediction aims to forecast the traffic volume (*e.g.,* inflow and outflow of each region at a given time), from past traffic observations. Recent advances have significantly boosted the research of traffic flow prediction with various deep learning techniques, *e.g.,* convolutional neural networks over region grids (Zhang, Zheng, and Qi 2017), graph neural networks for spatial dependency modeling (Zhang et al. 2021), and attention mechanism for spatial information aggregation (Zheng et al. 2020). Although significant efforts have been made to improve the traffic flow prediction results, existing models still face two key shortcomings.

The first limitation is the *lack of modeling spatial heterogeneity* exhibited with skewed traffic distributions across different regions. Taking Fig. 1(a) for example, A and B are two real-world regions in Beijing with different urban functions, namely the residential area and transportation hub. We can observe their quite different traffic flow distributions from Fig. 1(b). However, most existing models ignore such spatial heterogeneity and are easily biased towards popular regions with higher traffic volume, which make them insufficient to learn quality citywide traffic pattern representations. While some studies attempt to capture the heterogeneous flow distributions with multiple parameter sets over different regions (Pan et al. 2019b; Bai et al. 2020), the involved large parameter size may lead to the suboptimal issue over the skewed-distributed traffic data. Worse still, the high computational and memory cost of these methods make them infeasible to handle large-scale traffic data in practical urban scenarios. In addition, meta-learning has been used in recent approaches (Pan et al. 2019a; Ye et al. 2022) to con-

*Corresponding author: jywang@buaa.edu.cn

[1]The paper was done when Jiahao Ji was an intern at JD Intelligent Cities Research under the supervision of Junbo Zhang (msjunbozhang@outlook.com).

sider the difference of region traffic distributions. However, the effectiveness of those models largely relies on the collected handcrafted region spatial characteristics, *e.g.,* nearby points of interest and density of road networks, which limits the model representation generalization ability.

Furthermore, current traffic prediction methods model the temporal dynamics with a shared parameter space for all time periods, which can hardly precisely preserve the *temporal heterogeneity* in the latent embedding space. In real-life scenarios, traffic patterns of different regions vary over time, *e.g.,* from morning to evening, which results in the temporal heterogeneity as shown in Fig. 1(c). Nevertheless, the parameter space differentiation strategy adopted in (Song et al. 2020; Li and Zhu 2021) assumes that the temporal heterogeneity is static across the entire time periods, which is not always held, *e.g.,* evening traffic patterns can be significantly different for workdays and holidays shown in Fig. 1(c).

To effectively model both spatial and temporal heterogeneity, we present a novel Spatio-Temporal Self-Supervised Learning framework for predicting traffic flow. To encode spatial-temporal traffic patterns, our ST-SSL is built over a graph neural network which integrates temporal and spatial convolutions for information aggregation. To capture the spatial heterogeneity, we design a spatial self-supervised learning paradigm to augment the traffic flow graph at both data-level and structure-level, which is adaptive to the heterogeneous region traffic distributions. Then, the auxiliary self-supervision with a soft clustering paradigm is introduced to be aware of the diverse spatial patterns among different regions. To inject the temporal heterogeneity into our latent representation space, we empower ST-SSL to maintain dedicated representations of temporal traffic dynamics with temporal self-supervised learning paradigm. We summarize the key contributions of this work as follows:

- To our best of our knowledge, we are the first to propose a novel self-supervised learning framework to model spatial and temporal heterogeneity in traffic flow prediction. This paradigm may shed light on other practical spatio-temporal applications, such as air quality prediction.
- We propose an adaptive heterogeneity-aware data augmentation scheme over the graph-structured spatial-temporal graph against the noise perturbation.
- Two self-supervised learning tasks are incorporated to supplement the main traffic prediction task by enforcing the model discrimination ability with the awareness of both spatial and temporal traffic heterogeneity.
- Extensive experiments are conducted on four real-world public datasets to show the consistent performance superiority achieved by our ST-SSL across various settings.

2 Preliminaries

Definition 1 (Spatial Region)**.** *We partition a city into $N = I \times J$ disjoint geographical grids, in which each grid is considered as a spatial region $r_n (1 \leq n \leq N)$. We use $\mathcal{V} = \{r_1, \ldots, r_N\}$ to denote the spatial region set in a city.*

Definition 2 (Traffic Flow Graph (TFG))**.** *A traffic flow graph is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{A}, \mathcal{X}_{t-T:t})$, where $\mathcal{V}$ is the set of spatial regions (nodes) with the size of $|\mathcal{V}| = N$, and $\mathcal{E}$ is a set of edges connecting two spatially adjacent regions in $\mathcal{V}$. The adjacent matrix of our traffic flow graph is denoted as $\boldsymbol{A} \in \mathbb{R}^{N \times N}$. We represent the citywide traffic inflow and outflow data over previous T time steps with a traffic tensor $\mathcal{X}_{t-T:t} \in \mathbb{R}^{T \times N \times 2} = (\boldsymbol{X}_{t-T}, \ldots, \boldsymbol{X}_t)$. The traffic volume information of all regions $\mathcal{V}$ at the t-th time slot is denoted as $\boldsymbol{X}_t \in \mathbb{R}^{N \times 2}$.*

Problem Statement. Given the historical traffic flow graph $\mathcal{G}$ till the current time step, we aim to learn a predictive function which accurately estimates the traffic volume of all regions at the future time step $t+1$, *i.e.,* $\boldsymbol{X}_{t+1} \in \mathbb{R}^{N \times 2}$.

3 Methodology

This section elaborates on the technical details of our ST-SSL model with the overall architecture shown in Fig. 2.

3.1 Spatio-Temporal Encoder

We firstly propose a spatio-temporal (ST) encoder to jointly preserve the ST contextual information over the traffic flow graph, so as to jointly model the sequential patterns of traffic data across different time steps and the geographical correlations among spatial regions. Towards this end, we integrate the temporal convolutional component with the graph convolutional propagation network as the backbone for spatial-temporal relational representation.

For encoding the temporal traffic patterns, we adopt the 1-D causal convolution along the time dimension with a gated mechanism (Yu, Yin, and Zhu 2018). Specifically, our temporal convolution (TC) takes the traffic flow tensor as the input and outputs a time-aware embedding for each region:

$$(\boldsymbol{B}_{t-T_{out}}, \ldots, \boldsymbol{B}_t) = \mathrm{TC}\left(\boldsymbol{X}_{t-T}, \ldots, \boldsymbol{X}_t\right), \tag{1}$$

where $\boldsymbol{B}_t \in \mathbb{R}^{N \times D}$ denotes the region embedding matrix at the time step t. The n-th row $\boldsymbol{b}_{t,n} \in \mathbb{R}^D$ corresponds to the embedding of region r_n. Here, D denotes the embedding dimensionality. T_{out} is the length of the output embedding sequence after convolutional operations in TC encoder.

For capturing the region-wise spatial correlations, we design our spatial convolution (SC) encoder based on a graph-based message passing mechanism presented as follows:

$$\boldsymbol{E}_t = \mathrm{SC}\left(\boldsymbol{B}_t, \boldsymbol{A}\right). \tag{2}$$

$\boldsymbol{A}$ is the region adjacency matrix of $\mathcal{G}$. After our SC encoder, we can obtain the refined embeddings $(\boldsymbol{E}_{t-T_{out}}, \ldots, \boldsymbol{E}_t)$ of all regions by injecting the geographical context.

Our ST encoder is built with a "sandwich" block structure, in which TC $\rightarrow$ SC $\rightarrow$ TC is each individual block. By stacking multiple blocks, we can obtain a sequence of embedding matrix $(\boldsymbol{H}_{t-T'}, \ldots, \boldsymbol{H}_t)$ with the temporal dimension of T' after several convolutions. After ST encoder-based embedding propagation and aggregation, the temporal dimension T' reduces to zero and we generate the final embedding matrix $\boldsymbol{H} \in \mathbb{R}^{N \times D}$ for our ST encoder, in which each row $\boldsymbol{h}_n \in \mathbb{R}^D$ denotes the final embedding of region r_n.

In the next subsection, we will perform the adaptive augmentation over the $(\boldsymbol{B}_{t-T}, \ldots, \boldsymbol{B}_t)$ output from the first TC encoder layer (Sec 3.2), and self-supervised learning with the spatial-temporal heterogeneity modeling based on the final region embedding matrix $\boldsymbol{H}$ (Sec 3.3-Sec 3.4).

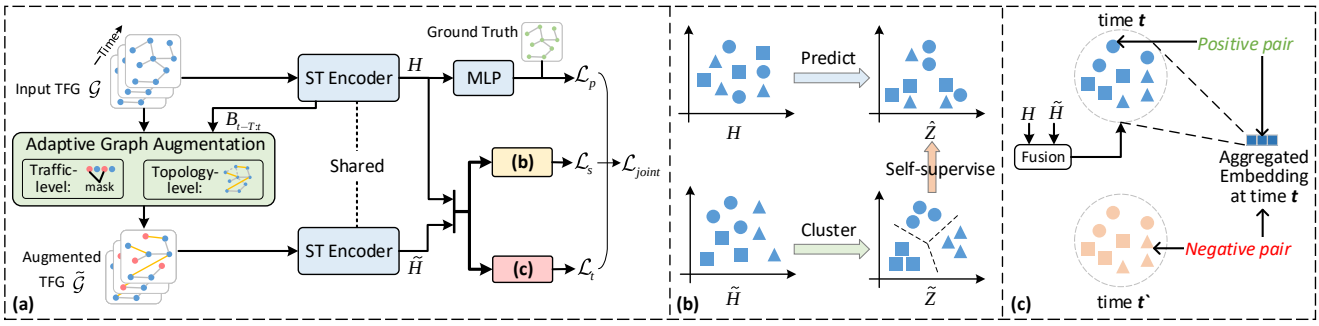


Figure 2: (a): The overall architecture of ST-SSL. (b): Spatial heterogeneity modeling. (c): Temporal heterogeneity modeling.

3.2 Adaptive Graph Augmentation on TFG

We devise two phases of graph augmentation schemes on TFG $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{A}, \mathcal{X}_{t-T:t})$ with traffic-level data augmentation and graph topology-level structure augmentation, which is adaptive to the learned heterogeneity-aware region dependencies in terms of their traffic regularities.

Region-wise Heterogeneity Measurement. For a region r_n, its embedding sequence $(\boldsymbol{b}_{t-T,n}, \ldots, \boldsymbol{b}_{t,n})$ within T time steps from rows of $(\boldsymbol{B}_{t-T}, \ldots, \boldsymbol{B}_t)$ is used to generate an overall embedding as:

$$\boldsymbol{u}_n = \sum_{\tau=t-T}^{t} p_{\tau,n} \cdot \boldsymbol{b}_{\tau,n}, \text{ where } p_{\tau,n} = \boldsymbol{b}_{\tau,n}^\top \cdot \boldsymbol{w}_0. \quad (3)$$

$\boldsymbol{u}_n$ is the aggregated representation over region r_n's embedding sequence across different time steps based on the derived aggregation weight $p_{\tau,n}$. Here, τ is the index of the time step range $(t-T, t)$. The aggregation weight $p_{\tau,n}$ reflects the relevance between the time step-specific traffic pattern $(\boldsymbol{b}_{\tau,n})$ and the overall traffic transitional regularities $(\boldsymbol{u}_n)$. $\boldsymbol{b}_{\tau,n}$ is region r_n's embedding at time step τ and $\boldsymbol{w}_0 \in \mathbb{R}^D$ is a learnable parameter vector for transformation.

In our ST-SSL model, we propose to estimate the heterogeneity degree between two regions, to be reflective of their traffic distribution difference over time below:

$$q_{m,n} = \frac{\boldsymbol{u}_m^\top \boldsymbol{u}_n}{\|\boldsymbol{u}_m\| \|\boldsymbol{u}_n\|}. \quad (4)$$

Note that a larger $q_{m,n}$ score indicates the higher traffic pattern dependencies between region r_m and r_n, thus resulting in the lower heterogeneity degree.

Heterogeneity-guided Data Augmentation. In our ST-SSL, we propose to perform data augmentation from both the traffic-level and graph topology-level elaborated below:

Traffic-level Augmentation. Inspired by the data augmentation strategy in (Zhu et al. 2021), we design an augmentation operator over the constructed traffic tensor $\mathcal{X}_{t-T:t}$, which is adaptive to the learned time-aware traffic pattern dependencies of each region. In particular, we aim to mask less relevant traffic volume at τ-th time step of region r_n against noise perturbation, based on a derived mask probability $\rho_{\tau,n}$ draw from a Bernoulli distribution *i.e.,* $\rho_{\tau,n} \sim \text{Bern}(1 - p_{\tau,n})$. The higher $\rho_{\tau,n}$ value indicates that region r_n's traffic volume $\boldsymbol{x}_{\tau,n}$ at τ-th time step is more likely to be masked, due to its lower relevance to the overall traffic regularities of region r_n. The augmented data with the traffic-level augmentation is denoted as $\tilde{\mathcal{X}}_{t-T:t}$.

Graph Topology-level Augmentation. In addition to the traffic-level augmentation, we propose to further perform the topology-level augmentation over the region traffic flow graph $\mathcal{G}$. By doing so, ST-SSL can not only debias the region connections with low inter-correlated traffic patterns, but also capture the long-range region dependencies with the global urban context. Towards this end, *i*) Given two spatially adjacent regions r_m and r_n, their connection edge $(r_m, r_n) \in \mathcal{E}$ will be masked if they are not highly dependent in terms of their traffic regularities, measured by the high heterogeneity degree $q_{m,n}$. The mask probability $\rho_{m,n}$ is drawn from a Bernoulli distribution *i.e.,* $\rho_{m,n} \sim \text{Bern}(1-q_{m,n})$. *ii*) Given two non-adjacent regions, the low heterogeneity degree $q_{m,n}$ will result in adding an edge between r_m and r_n based on the masking probability drawn from a Bernoulli distribution, $\text{Bern}(q_{m,n})$ similarly.

After two augmentation phases, we obtain the augmented TFG $\tilde{\mathcal{G}} = \left(\mathcal{V}, \tilde{\mathcal{E}}, \tilde{\boldsymbol{A}}, \tilde{\mathcal{X}}_{t-T:t}\right)$, with the debiased traffic volume input $\tilde{\mathcal{X}}_{t-T:t}$ (traffic-level augmentation) and structure denoising $\tilde{\mathcal{E}}, \tilde{\boldsymbol{A}}$ (graph topology-level augmentation).

3.3 SSL for Spatial Heterogeneity Modeling

Given the heterogeneity-aware augmented TFG, we aim to enable the region embeddings to effectively preserve the spatial heterogeneity with auxiliary self-supervised signals.

To achieve this goal, we design a soft clustering-based self-supervised learning (SSL) task over regions, to map them into multiple latent representation spaces corresponding to diverse urban region functionalities (*e.g.,* residential zone, shopping mall, transportation hub). Specifically, we generate K cluster embeddings $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K\}$ (indexed by k) as latent factors for region clustering. Formally, the clustering process is performed with $\tilde{z}_{n,k} = \boldsymbol{c}_k^\top \tilde{\boldsymbol{h}}_n$. Here, $\tilde{\boldsymbol{h}}_n \in \mathbb{R}^D$ is the region embedding of region r_n encoded from the augmented TFG $\tilde{\mathcal{G}}$. $\tilde{z}_{n,k}$ represents the estimated relevance score between region r_n's embedding and the embedding $\boldsymbol{c}_k$ of the k-th cluster. Afterwards, the cluster assignment of region r_n is generated with $\tilde{\boldsymbol{z}}_n = (\tilde{z}_{n,1}, \ldots, \tilde{z}_{n,K})^\top$.

To provide self-supervised signals based on the heterogeneity-aware soft clustering paradigm for augmentation, the auxiliary learning task is designed to predict

the cluster assignment using the region embedding $\boldsymbol{h}_n$ encoded from the original TFG $\mathcal{G}$ as: $\hat{z}_{n,k} = \boldsymbol{c}_k^\top \boldsymbol{h}_n$, where $\hat{\boldsymbol{z}}_{n,k}$ is the predicted assignment score for $\tilde{\boldsymbol{z}}_{n,k}$. The self-supervised augmented task is optimized as follows:

$$\ell(\boldsymbol{h}_n, \tilde{\boldsymbol{z}}_n) = -\sum_k \tilde{z}_{n,k} \log \frac{\exp\left(\hat{z}_{n,k}/\gamma\right)}{\sum_j \exp\left(\hat{z}_{n,j}/\gamma\right)}, \tag{5}$$

where γ is the temperature parameter to control the smoothing degree of softmax output. The overall self-supervised objective over all regions is defined as follows:

$$\mathcal{L}_s = \sum_{n=1}^{N} \ell(\boldsymbol{h}_n, \tilde{\boldsymbol{z}}_n). \tag{6}$$

By incorporating the supervision on $\boldsymbol{h}_n$ with the heterogeneity-aware region cluster assignment $\tilde{\boldsymbol{z}}_n$, we make the region embedding $\boldsymbol{h}_n$ to be reflective of spatial heterogeneity within the global urban space.

Distribution Regularization for Region Clustering. In our heterogeneity-aware region clustering paradigm, we generate the cluster assignment matrix $\tilde{\boldsymbol{Z}} = (\tilde{\boldsymbol{z}}_1, \ldots, \tilde{\boldsymbol{z}}_N)^\top \in \mathbb{R}^{N\times K}$ as self-supervised signals for generative data augmentation. However, two issues need to be addressed to fit the true distribution of regional characteristics in urban space: i) Since $\tilde{\boldsymbol{Z}}$ is produced by matrix production, there is no guarantee that each region's cluster assignment sums up to 1, *i.e.,* $\tilde{\boldsymbol{Z}}\mathbf{1}_K = \mathbf{1}_N$, where $\mathbf{1}_N$ denotes an N-dimensional vector of all ones. ii) To avoid the trivial solution that every region has the same assignment, we employ the principle of maximum entropy, *i.e.,* $\tilde{\boldsymbol{Z}}^\top \mathbf{1}_N = \frac{N}{K}\mathbf{1}_K$. This encourages all regions to be equally partitioned by the clusters. To tackle these two issues, we define a feasible solution set as:

$$\tilde{\mathcal{Z}} = \left\{ \tilde{\boldsymbol{Z}} \in \mathbb{R}_+^{N\times K} \middle| \tilde{\boldsymbol{Z}}\mathbf{1}_K = \mathbf{1}_N, \tilde{\boldsymbol{Z}}^\top \mathbf{1}_N = \frac{N}{K}\mathbf{1}_K \right\}. \tag{7}$$

For any assignment $\tilde{\boldsymbol{Z}} \in \tilde{\mathcal{Z}}$, we can use it to map the embedding matrix $\tilde{\boldsymbol{H}} = (\tilde{\boldsymbol{h}}_1, \ldots, \tilde{\boldsymbol{h}}_N)^\top \in \mathbb{R}^{N\times D}$ into the cluster matrix $\boldsymbol{C} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K)^\top \in \mathbb{R}^{K\times D}$. Thus, we search for the optimal solution by maximizing the similarity between the embeddings and the clusters, *i.e.,*

$$\max_{\tilde{\boldsymbol{Z}}\in\mathcal{Z}} \operatorname{tr}\left(\tilde{\boldsymbol{Z}}\boldsymbol{C}\tilde{\boldsymbol{H}}^\top\right) + \epsilon H(\tilde{\boldsymbol{Z}}), \tag{8}$$

where $\operatorname{tr}(\cdot)$ is the trace operator that sums elements on the main diagonal of a square matrix, $H(\tilde{\boldsymbol{Z}})$ is the entropy function defined as $-\sum_{n,k} \tilde{z}_{n,k} \log \tilde{z}_{n,k}$, and ϵ is a parameter that controls the smoothness of the assignment. Finally, the original assignment in Eq. (6) is replaced with the optimal solution. Refer to the Appendix for the solution procedure.

3.4 SSL for Temporal Heterogeneity Modeling

In this component, we further design a self-supervised learning (SSL) task to inject the temporal heterogeneity into time-aware region embeddings, by enforcing the divergence among time step-specific traffic pattern representations.

Specifically, we firstly fuse the encoded time-aware region embeddings from both the original and augmented TFGs:

$$\boldsymbol{v}_{t,n} = \boldsymbol{w_1} \odot \boldsymbol{h}_{t,n} + \boldsymbol{w_2} \odot \tilde{\boldsymbol{h}}_{t,n}, \tag{9}$$

where $\odot$ is the element-wise product. $\boldsymbol{w}_1, \boldsymbol{w}_2$ are learnable parameters. After that, we generate the city-level representation $\boldsymbol{s}_t$ at the time step t through aggregating embeddings of all regions (σ is the sigmoid function):

$$\boldsymbol{s}_t = \sigma\left(\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{v}_{t,n}\right). \tag{10}$$

To enhance the representation discrimination ability among different time steps, we treat the region-level and city-level embeddings $(\boldsymbol{v}_{t,n}, \boldsymbol{s}_t)$ from the same time step as the positive pairs in our SSL task, and the embeddings from different time steps as negative pairs. With this design, the auxiliary supervision of positive pairs will encourage the consistency of time-specific citywide traffic trends (*e.g.,* rush hours, weather factors), while the negative pairs help in capturing the temporal heterogeneity across different time steps. Formally, the temporal heterogeneity-enhanced SSL task is optimized with the following loss with cross-entropy metric:

$$\mathcal{L}_t = -\left(\sum_{n=1}^{N} \log g\left(\boldsymbol{v}_{t,n}, \boldsymbol{s}_t\right) + \sum_{n=1}^{N} \log\left(1 - g\left(\boldsymbol{v}_{t',n}, \boldsymbol{s}_t\right)\right)\right), \tag{11}$$

where t and t' denote two different time steps. g is a criterion function defined as $g\left(\boldsymbol{v}_{t,n}, \boldsymbol{s}_t\right) = \sigma\left(\boldsymbol{v}_{t,n}^\top \boldsymbol{W}_3 \boldsymbol{s}_t\right)$. $\boldsymbol{W}_3 \in \mathbb{R}^{N\times N}$ is the learnable transformation matrix.

3.5 Model Training

In the learning process of our ST-SSL, we feed the embedding $\boldsymbol{h}_n \in \boldsymbol{H}$ of each region r_n into an MLP structure to enable the traffic flow prediction at the future time step $t+1$ as:

$$\hat{\boldsymbol{x}}_{t+1,n} = \mathrm{MLP}(\boldsymbol{h}_n), \tag{12}$$

where $\hat{\boldsymbol{x}}_{t+1,n}$ is the predicted result. The model is optimized by minimizing the loss function below:

$$\mathcal{L}_p = \sum_{n=1}^{N} \lambda \left| x_{t+1,n}^{(0)} - \hat{x}_{t+1,n}^{(0)} \right| + (1-\lambda) \left| x_{t+1,n}^{(1)} - \hat{x}_{t+1,n}^{(1)} \right|, \tag{13}$$

where $x_{t+1,n}^{(0)}, x_{t+1,n}^{(1)}$ denote the ground truth of inflow and outflow respectively. λ is a parameter to balance the influence of each type of traffic flow.

Finally, we obtain the overall loss by incorporating the self-supervised spatial and temporal heterogeneity modeling losses in Eq. (6) and (11) into the joint learning objective:

$$\mathcal{L}_{joint} = \mathcal{L}_p + \mathcal{L}_s + \mathcal{L}_t. \tag{14}$$

Our model can be trained via the back-propagation algorithm. The entire training procedure can be summarized into four stages: i) given a TFG $\mathcal{G}$, we generate a region embedding matrix $\boldsymbol{H}$ by the ST encoder. ii) Meanwhile, we perform adaptive augmentation to refine $\mathcal{G}$ as $\tilde{\mathcal{G}}$, which is fed into the shared ST encoder to output $\tilde{\boldsymbol{H}}$. iii) By using $\boldsymbol{H}$ and $\tilde{\boldsymbol{H}}$, we calculate the losses $\mathcal{L}_s$, $\mathcal{L}_t$, and $\mathcal{L}_p$ that are used to produce the joint loss $\mathcal{L}_{joint}$. iv). We employ the back-propagation algorithm to train ST-SSL until $\mathcal{L}_{joint}$ converges.

4 Experiments

In this section, we evaluate the performance of ST-SSL on a series of experiments over several real-world datasets, which are summarized to answer the following research questions:

• **RQ1**: How is the overall traffic prediction performance of ST-SSL as compared to various baselines?

• **RQ2**: How do designed different sub-modules contribute to the model performance?

• **RQ3**: How does ST-SSL perform with regard to heterogeneous spatial regions and different time periods?

• **RQ4**: How do the augmented graph and learned representations benefit the model?

4.1 Experimental Settings

Data Description. We evaluate our model on two types of public real-world traffic datasets summarized in Tab. 1.

The first kind is about bike rental records in New York City. **NYCBike1** (Zhang, Zheng, and Qi 2017) spans from 04/01/2014 to 09/30/2014, and **NYCBike2** (Yao et al. 2019) spans from 07/01/2016 to 08/29/2016. They are all measured every 30 minutes. The second kind is about taxi GPS trajectories. **NYCTaxi** (Yao et al. 2019) spans from 01/01/2015 to 03/01/2015. Its time interval is half an hour. **BJTaxi** (Zhang, Zheng, and Qi 2017), collected in Beijing, spans from 03/01/2015 to 06/30/2015 on an hourly basis.

For all datasets, previous 2-hour flows as well as previous 3-day flows around the predicted time are used to predict the flows for the next time step. We use a sliding window strategy to generate samples, and then split each dataset into the training, validation, and test sets with a ratio of 7:1:2.

Evaluation Metrics & Baselines. In our experiments, two common metrics are used for evaluation: Mean Average Error (MAE) and Mean Average Percentage Error (MAPE). We compare our proposed ST-SSL with 8 baselines that fall into three categories.

Traditional Time Series Prediction Approaches:

• **ARIMA** (Kumar and Vanajakshi 2015): it is a classical time series prediction model.

• **SVR** (Castro-Neto et al. 2009): it is a regression model widely used for time series analysis.

Spatial-Temporal Prediction Methods:

• **ST-ResNet** (Zhang, Zheng, and Qi 2017): it is a convolution-based model that constructs multiple traffic time series to capture the temporal dependencies and utilizes residual convolution to model the spatial correlations.

• **STGCN** (Yu, Yin, and Zhu 2018): it is a graph convolution-based model that combines 1D convolution to capture spatial and temporal correlations, respectively.

• **GMAN** (Zheng et al. 2020): it is an attention-based model that adopts an encoder-decoder architecture for traffic flow prediction.

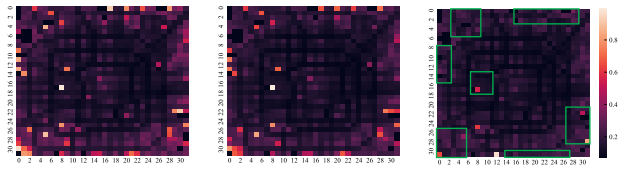
Spatial-Temporal Methods Considering Heterogeneity:

• **AGCRN** (Bai et al. 2020): it enhances the traditional graph convolution by adaptive modules and combines them into recurrent networks to capture spatial-temporal correlations.

• **STSGCN** (Song et al. 2020): it captures the complex localized spatial-temporal correlations through a spatial-temporal synchronous modeling mechanism.

Data type	Bike rental		Taxi GPS	
Dataset	NYCBike1	NYCBike2	NYCTaxi	BJTaxi
Time interval	1 hour	30 min	30 min	30 min
# regions	16×8	10×20	10×20	32×32
# taxis/bikes	6.8k+	2.6m+	22m+	34k+

Table 1: Statistics of Datasets.



(a) ST-ResNet (b) AGCRN (c) ST-SSL

Figure 3: Visualization of traffic prediction errors.

• **STFGNN**: (Li and Zhu 2021) it integrates with STFGN module and a novel gated CNN module, and captures hidden spatial dependencies by a data-driven graph and its further fusion with given spatial graphs.

Methods in the last category model the traffic heterogeneity by using multiple parameter spaces.

Parameter Settings. The ST-SSL is implemented with PyTorch. The embedding dimension D is set as 64. Both the temporal and spatial convolution kernel sizes of ST encoder are set to 3. The perturbation ratios for both traffic-level and topology-level augmentations are set as 0.1. The training phase is performed using the Adam optimizer and a batch size of 32. The experiments of baselines are performed with their released codes on the LibCity (Wang et al. 2021) platform.

4.2 Performance Comparison (RQ1)

Table 2 shows the comparison results of all methods. We run all deep learning models with 5 different seeds and report the average performance and their standard deviations.

Performance Superiority of ST-SSL. According to Student's t-test at level 0.01, our ST-SSL significantly outperforms other competing baselines with regard to both metrics over all datasets. This demonstrates the effectiveness of ST-SSL in jointly modeling the spatial and temporal heterogeneity in a self-supervised manner. Fig. 3 visualizes the prediction error ($|\hat{x}_n - x_n|/x_n$) of ST-SSL and two best performed baselines on BJTaxi dataset, where a brighter pixel means a larger error. The superiority of our model can still be observed, which is consistent with the quantitative results in Table 2. Interestingly, ST-SSL exhibits a significant improvement in the suburban areas (green boxes in Fig. 3), which justifies the effectiveness of spatial heterogeneity modeling that transfers information among global similar regions.

Performance Comparison between Baselines. Spatio-temporal prediction methods outperform time series approaches in most cases, which suggests the necessity to capture spatial dependencies. The methods that take into account the heterogeneity of traffic data usually perform bet-

Dataset	Metric	Type	ARIMA	SVR	ST-ResNet	STGCN	GMAN	AGCRN	STSGCN	STFGNN	ST-SSL
NYCBike1	MAE	In	10.66	7.27	5.53±0.06	5.33±0.02	6.77±3.42	5.17±0.03	5.81±0.04	6.53±0.10	**4.94±0.02**
		Out	11.33	7.98	5.74±0.07	5.59±0.03	7.17±3.61	5.47±0.03	6.10±0.04	6.79±0.08	**5.26±0.02**
	MAPE	In	33.05	25.39	25.46±0.20	26.92±0.08	31.72±12.29	25.59±0.22	26.51±0.32	32.14±0.23	**23.69±0.11**
		Out	35.03	27.42	26.36±0.50	27.69±0.14	34.74±17.04	26.63±0.30	27.56±0.39	32.88±0.19	**24.60±0.27**
NYCBike2	MAE	In	8.91	12.82	5.63±0.14	5.21±0.02	5.24±0.13	5.18±0.03	5.25±0.03	5.80±0.10	**5.04±0.03**
		Out	8.70	11.48	5.26±0.08	4.92±0.02	4.97±0.14	4.79±0.04	4.94±0.05	5.51±0.11	**4.71±0.02**
	MAPE	In	28.86	46.52	32.17±0.85	27.73±0.16	27.38±1.13	27.14±0.14	29.26±0.13	30.73±0.49	**22.54±0.10**
		Out	28.22	41.91	30.48±0.86	26.83±0.21	26.75±1.14	26.17±0.22	28.02±0.23	29.98±0.46	**21.17±0.13**
NYCTaxi	MAE	In	20.86	52.16	13.48±0.14	13.12±0.04	15.09±0.61	12.13±0.11	13.69±0.11	16.25±0.38	**11.99±0.12**
		Out	16.80	41.71	10.78±0.25	10.35±0.03	12.06±0.39	9.87±0.04	10.75±0.17	12.47±0.25	**9.78±0.09**
	MAPE	In	21.49	65.10	24.83±0.55	21.01±0.18	22.73±1.20	18.78±0.04	22.91±0.44	24.01±0.30	**16.38±0.10**
		Out	21.23	64.06	24.42±0.52	20.78±0.16	21.97±0.86	18.41±0.21	22.37±0.16	23.28±0.47	**16.86±0.23**
BJTaxi	MAE	In	21.48	52.77	12.12±0.11	12.34±0.09	13.13±0.43	12.30±0.06	12.72±0.03	13.83±0.04	**11.31±0.03**
		Out	21.60	52.74	12.16±0.12	12.41±0.08	13.20±0.43	12.38±0.06	12.79±0.03	13.89±0.04	**11.40±0.02**
	MAPE	In	23.12	65.51	15.50±0.26	16.66±0.21	18.67±0.99	15.61±0.15	17.22±0.17	19.29±0.07	**15.03±0.13**
		Out	20.67	65.51	15.57±0.26	16.76±0.22	18.84±1.04	15.75±0.15	17.35±0.17	19.41±0.07	**15.19±0.15**

Table 2: Model comparison on four datasets in terms of MAE and MAPE (%). In and Out represent the inflow and outflow.

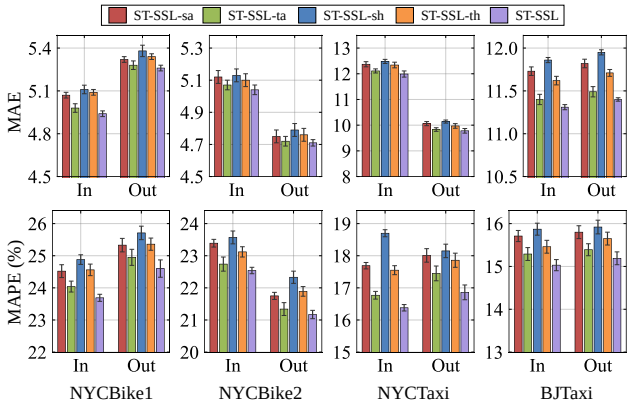
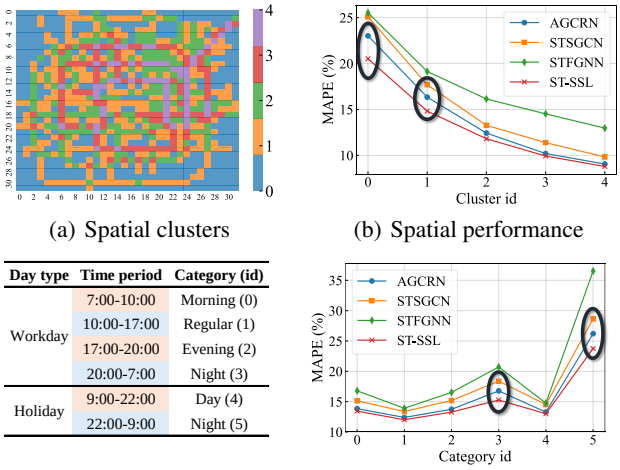


Figure 4: Ablation study of our proposed ST-SSL.



(a) Spatial clusters

(b) Spatial performance

Day type	Time period	Category (id)
Workday	7:00-10:00	Morning (0)
	10:00-17:00	Regular (1)
	17:00-20:00	Evening (2)
	20:00-7:00	Night (3)
Holiday	9:00-22:00	Day (4)
	22:00-9:00	Night (5)

(c) Temporal categories

(d) Temporal performance

Figure 5: Prediction performance with regard to heterogeneous spatial regions and different time periods.

ter than those that use shared parameters across different regions and time periods, indicating the rationality of learning spatial and temporal heterogeneity in traffic prediction.

4.3 Ablation Study (RQ2)

To analyze the effects of sub-modules in our ST-SSL framework, we perform ablation studies with five variants:

- **ST-SSL-sa**: This variant replaces heterogeneity-guided structure augmentation on graph topology with random edge removal and addition augmentations.
- **ST-SSL-ta**: ST-SSL replaces heterogeneity-guided traffic-level augmentation with random traffic volume masking augmentations.
- **ST-SSL-sh**: ST-SSL without spatial heterogeneity modeling.
- **ST-SSL-th**: ST-SSL without temporal heterogeneity modeling.

The results are present in Fig. 4. We can observe that ST-SSL beats the variants with random augmentation, indicating the effectiveness of adaptive heterogeneity-guided data augmentation at both traffic-level and graph structure-level. Moreover, ST-SSL consistently outperforms ST-SSL-sh and ST-SSL-th, which justifies the necessity to jointly model the spatial and temporal heterogeneity. In summary, each designed sub-module has a positive effect on performance improvement.

4.4 Robustness Analysis (RQ3)

To explore the robustness of our ST-SSL, we perform traffic prediction for spatial regions with heterogeneous data distributions and time periods with different patterns on BJTaxi. Specifically, we cluster regions by using traffic data statistics, *i.e.,* $(mean, median, standard\ deviation)$ of their historical traffic flow. As shown in Fig. 5(a), regions with

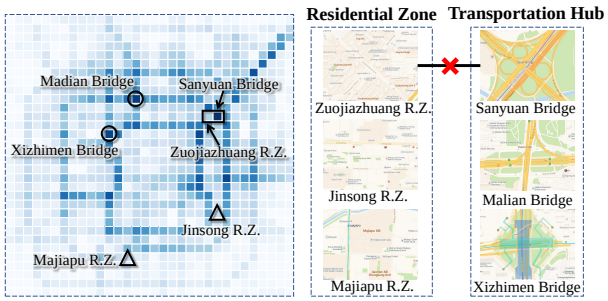


Figure 6: Case study on the adaptive graph augmentation.

smaller cluster id (next to the color bar) are usually located in suburbs that are less popular and thus have lower traffic. Fig. 5(b) exhibits the prediction performance for different clusters. Our ST-SSL surpasses other baselines by a significant margin, particularly for less popular regions (marked by black circles), which is consistent with results in Fig. 3. This also verifies the robustness of ST-SSL to accurately predicts traffic flows of different types of spatial regions.

For temporal heterogeneity, according to urban traffic rhythms (Wang et al. 2019a), we partition a workday into four time periods and a holiday (weekend included) into two time periods, whose categories are given in Fig. 5(c). Fig. 5(d) presents the evaluation performance. Our ST-SSL beats the baselines in terms of every category. Furthermore, ST-SSL shows a significant improvement in categories 3 and 5 that denote the nighttime of workdays and holidays. During these times, traffic flow data are typically sparse, making it difficult for baselines to produce accurate predictions. ST-SSL can handle this situation because we inject the temporal heterogeneity into the time-aware region embeddings.

4.5 Qualitative Study (RQ4)

In Fig. 6, we investigate the heterogeneity-guided graph topology-level augmentation on BJTaxi. Our augmentation method adaptively removes connections between adjacent regions with heterogeneous traffic patterns, *i.e.,* Zuojiazhuang Residential Zone and Sanyuan Bridge (a transportation hub). Meanwhile, it builds connections between distant regions with similar latent urban function, *e.g.,* Xizhimen Bridge and Sanyuan Bridge that are both transportation hubs. In this way, our ST-SSL can not only debias the region connections with low inter-correlated traffic patterns, but also capture the long-range region dependencies with the global urban context.

To further explore why the embeddings obtained by ST-SSL can deliver more accurate traffic prediction than AGCRN, we visualize them on BJTaxi by t-SNE (Van der Maaten and Hinton 2008). We plot the learned embeddings of all regions with ground truth classes the same as Fig. 5(a). As shown in Fig. 7, samples in the same class are more compact and those of different classes are significantly better separated for ST-SSL. This enables ST-SSL to be aware of spatial heterogeneity and transfer information between regions in the same class, which facilitates predictions.

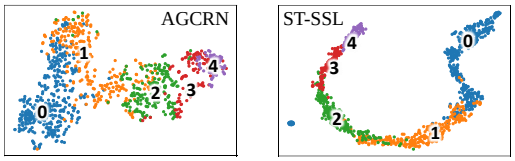


Figure 7: t-SNE visualization of embeddings on BJTaxi.

5 Related Work

Deep Learning for Traffic Prediction. Many efforts have been devoted to developing traffic prediction techniques based on various neural networks. RNN (Wang et al. 2019b; Ji et al. 2020) and 1D CNN (Wang et al. 2022, 2016) are applied to capture the temporal dependencies in traffic series. CNN (Zhang, Zheng, and Qi 2017; Yao et al. 2019), GNN (Zhang et al. 2020; Ji et al. 2022), and attention mechanism (Zheng et al. 2020) are introduced to incorporate the spatial information. However, most of them neglect the spatio-temporal heterogeneity problem. Recently, some works model the heterogeneity by using multiple models (Yuan, Zhou, and Yang 2018) or multiple sets of parameters (Bai et al. 2020; Li and Zhu 2021), and some use meta learning to generate different weights based on static features of different regions (Pan et al. 2019a; Ye et al. 2022). However, these methods either introduce a number of parameters that may cause an overfitting problem or require external data that may be not available. To overcome these limitations, we incorporate self-supervised learning into traffic prediction to explore spatial and temporal heterogeneity.

Self-Supervised Learning for Representation Learning. Self-supervised learning aims to extract useful information from input data to improve the representation quality (Hendrycks et al. 2019). The general paradigm is to augment the input data and then design pretext tasks as pseudo-labels for representation learning. It has achieved great success with text (Kenton and Toutanova 2019), image (Chen et al. 2020), and audio data (Oord, Li, and Vinyals 2018). Motivated by these works, we develop an adaptive data augmentation method for spatio-temporal graph data and introduce two pretext tasks to learn representations that are robust to spatio-temporal heterogeneity, which has not been well explored in existing traffic flow prediction methods.

6 Conclusion and Future Work

This work investigated the traffic prediction problem by proposing a novel spatio-temporal self-supervised learning (ST-SSL) framework. Specifically, we integrated temporal and spatial convolutions to encode spatial-temporal traffic patterns. Then, we devised i) a spatial self-supervised learning paradigm that consists of an adaptive graph augmentation and a clustering-based generative task, and ii) a temporal self-supervised learning paradigm that relies on a time-aware contrastive task, to supplement the main traffic flow prediction task with spatial and temporal heterogeneity-aware self-supervised signals. Comprehensive experiments on four traffic flow datasets demonstrated the robustness of ST-SSL. The future work lies in extending our spatial-temporal SSL framework to a model-agnostic paradigm.

Acknowledgments

This work was supported by the National Key R&D Program of China (2019YFB2101804). Prof. Wang's work was supported by the National Natural Science Foundation of China (No. 72222022, 82161148011, 72171013), the Fundamental Research Funds for the Central Universities (YWF-22-L-838) and the DiDi Gaia Collaborative Research Funds. Dr. Zhang's work was supported by the National Natural Science Foundation of China (No. 62172034) and the Beijing Nova Program (Z201100006820053).

References

Bai, L.; Yao, L.; Li, C.; Wang, X.; and Wang, C. 2020. Adaptive graph convolutional recurrent network for traffic forecasting. *NeurIPS*, 33: 17804–17815.

Castro-Neto, M.; Jeong, Y.-S.; Jeong, M.-K.; and Han, L. D. 2009. Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions. *Expert Systems with Applications*, 36(3): 6164–6173.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607.

Hendrycks, D.; Mazeika, M.; Kadavath, S.; and Song, D. 2019. Using self-supervised learning can improve model robustness and uncertainty. *NeurIPS*, 32.

Ji, J.; Wang, J.; Jiang, Z.; Jiang, J.; and Zhang, H. 2022. STDEN: Towards physics-guided neural networks for traffic flow prediction. In *AAAI*, volume 36, 4048–4056.

Ji, J.; Wang, J.; Jiang, Z.; Ma, J.; and Zhang, H. 2020. Interpretable spatiotemporal deep learning model for traffic flow prediction based on potential energy fields. In *ICDM*, 1076–1081.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186.

Kumar, S. V.; and Vanajakshi, L. 2015. Short-term traffic flow prediction using seasonal ARIMA model with limited input data. *European Transport Research Review*, 7(3): 1–9.

Li, M.; and Zhu, Z. 2021. Spatial-temporal fusion graph neural networks for traffic flow forecasting. In *AAAI*, volume 35, 4189–4196.

Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Pan, Z.; Liang, Y.; Wang, W.; Yu, Y.; Zheng, Y.; and Zhang, J. 2019a. Urban traffic prediction from spatio-temporal data using deep meta learning. In *ACM SIGKDD*, 1720–1730.

Pan, Z.; Wang, Z.; Wang, W.; Yu, Y.; Zhang, J.; and Zheng, Y. 2019b. Matrix factorization for spatio-temporal neural networks with applications to urban flow prediction. In *CIKM*, 2683–2691.

Song, C.; Lin, Y.; Guo, S.; and Wan, H. 2020. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *AAAI*, volume 34, 914–921.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, J.; Gu, Q.; Wu, J.; Liu, G.; and Xiong, Z. 2016. Traffic speed prediction and congestion source exploration: A deep learning method. In *ICDM*, 499–508.

Wang, J.; Ji, J.; Jiang, Z.; and Sun, L. 2022. Traffic Flow Prediction Based on Spatiotemporal Potential Energy Fields. *IEEE Transactions on Knowledge and Data Engineering*, 1–14.

Wang, J.; Jiang, J.; Jiang, W.; Li, C.; and Zhao, W. X. 2021. LibCity: An open library for traffic prediction. In *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*, 145–148.

Wang, J.; Wu, J.; Wang, Z.; Gao, F.; and Xiong, Z. 2019a. Understanding urban dynamics via context-aware tensor factorization with neighboring regularization. *IEEE Transactions on Knowledge Data Engineering*, 32(11): 2269–2283.

Wang, J.; Wu, N.; Zhao, W. X.; Peng, F.; and Lin, X. 2019b. Empowering A* search algorithms with neural networks for personalized route recommendation. In *ACM SIGKDD*, 539–547.

Yao, H.; Tang, X.; Wei, H.; Zheng, G.; and Li, Z. 2019. Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction. In *AAAI*, volume 33, 5668–5675.

Ye, X.; Fang, S.; Sun, F.; Zhang, C.; and Xiang, S. 2022. Meta graph transformer: A novel framework for spatial-temporal traffic prediction. *Neurocomputing*, 491: 544–563.

Yu, B.; Yin, H.; and Zhu, Z. 2018. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In *IJCAI*, 3634–3640.

Yuan, Z.; Zhou, X.; and Yang, T. 2018. Hetero-ConvLSTM: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In *ACM SIGKDD*, 984–992.

Zhang, J.; Zheng, Y.; and Qi, D. 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *AAAI*, volume 31, 1655–1661.

Zhang, X.; Huang, C.; Xu, Y.; and Xia, L. 2020. Spatial-temporal convolutional graph attention networks for citywide traffic flow forecasting. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, 1853–1862.

Zhang, X.; Huang, C.; Xu, Y.; Xia, L.; Dai, P.; Bo, L.; Zhang, J.; and Zheng, Y. 2021. Traffic flow forecasting with spatial-temporal graph diffusion network. In *AAAI*, volume 35, 15008–15015.

Zheng, C.; Fan, X.; Wang, C.; and Qi, J. 2020. GMAN: A graph multi-attention network for traffic prediction. In *AAAI*, volume 34, 1234–1241.

Zhu, Y.; Xu, Y.; Yu, F.; Liu, Q.; Wu, S.; and Wang, L. 2021. Graph contrastive learning with adaptive augmentation. In *Proceedings of the Web Conference 2021*, 2069–2080.