# Traffic Flow Forecasting with Spatial-Temporal Graph Diffusion Network

**Xiyue Zhang**[1], **Chao Huang**[2*], **Yong Xu**[1,3,4], **Lianghao Xia**[1], **Peng Dai**[2]
**Liefeng Bo**[2], **Junbo Zhang**[5,6], **Yu Zheng**[5,6]
South China University of Technology[1], China
JD Finance America Corporation[2], USA
Communication and Computer Network Laboratory of Guangdong[3], China
Peng Cheng Laboratory[4], China
JD Intelligent Cities Research[5], China, JD Intelligent Cities Business Unit[6], JD Digits, China
{zhang.xiyue,cslianghao.xia}@mail.scut.edu.cn, yxu@scut.edu.cn, chaohuang75@gmail.com
{peng.dai,liefeng.bo}@jd.com, {msjunbozhang,msyuzheng}@outlook.com

## Abstract

Accurate forecasting of citywide traffic flow has been playing critical role in a variety of spatial-temporal mining applications, such as intelligent traffic control and public risk assessment. While previous work has made significant efforts to learn traffic temporal dynamics and spatial dependencies, two key limitations exist in current models: i) Most of these methods solely focus on neighboring spatial correlations among adjacent regions, and ignore the global geographical contextual information; ii) These methods fail to encode the complex traffic transition regularities exhibited with time-dependent and multi-resolution in nature. To tackle these challenges, we propose a new traffic flow prediction framework–**S**patial-**T**emporal **G**raph **D**iffusion **N**etwork. In particular, ST-GDN is a hierarchically structured graph neural architecture which learns not only the local region-wise geographical dependencies, but also the spatial semantics from a global perspective. Furthermore, a multi-scale attention network is developed to empower ST-GDN with the capability of capturing multi-level temporal dynamics. Experiments on four real-life traffic datasets demonstrate that ST-GDN outperforms different types of state-of-the-art baselines. Source codes are available at *https://github.com/jillbetty001/ST-GDN*.

## Introduction

Accurate forecasting of traffic flow across different geographical regions in a city, have played a critical role in smart transformation systems, such as intelligent transportation (Wei et al. 2018; Huang et al. 2020) and public risk assessment (Gao et al. 2019; Huang et al. 2018). For example, in disaster control, by predicting future traffic volume, local governments and communities is able to design better transportation scheduling and mobility management strategies, to mitigate the tragedies caused by the crowd flow (Zhao et al. 2017). In general, the objective of traffic prediction is to forecast the traffic volume (*e.g.*, inflow and outflow of each region), from past traffic observations (Diao et al. 2019).

Inspired by the advancement of deep learning techniques, many efforts have been devoted to developing traffic prediction methods with various neural network architecture for spatial-temporal pattern modeling. Inspired by the sequence learning paradigm, recent neural networks have been utilized to model temporal effects of traffic variations (Liu et al. 2016; Yu et al. 2017a). To make use of spatial features, some research work propose to adopt convolutional neural network to model correlations between adjacent regions (Zhang, Zheng, and Qi 2017), along with using recurrent neural layers on temporal dimension (Yao et al. 2018). Although both spatial and temporal correlations have been considered in existing methods, we identify three significant challenges that have not been addressed well.

In real-life scenarios, traffic flow pattern is often complex and multi-periodic (Zhang, Zheng, and Qi 2017; Deng et al. 2016), as each individual time resolution view (*e.g.*, hourly, daily, weekly) reflects traffic dynamics from different temporal dimensions. The captured temporal patterns are often complementary with each other (Wu et al. 2018). Hence, learning robust representations of traffic variation patterns requires the collaboration of multiple views with different time resolutions. While recurrent networks have achieved good performance on various spatial-temporal sequence prediction tasks, they can only be effective for short-term, smooth dynamics and hardly make predictions over high-order multi-dimensional time horizons (Yu et al. 2017b).

Most current forecasting approaches merely focus on modeling nearby geographical correlations (Yao et al. 2018; Zhang, Zheng, and Qi 2017), while ignoring the cross-region inter-dependencies under a global context. For example, two geographical areas with similar urban functions (*e.g.*, shopping zone or transportation hub) can be correlated in terms of their traffic distribution, although they are not spatially adjacent or even far away from each other (Shen et al. 2018; Wang and Li 2017). Hence, the learned region-wise relational structures without the global-level traffic transition information, are insufficient to distill not only local geographical dependencies, but also semantic relations across regions, which leads to suboptimal predictions.

To tackle the above challenges, we propose a new predictive framework **S**patial-**T**emporal **G**raph **D**iffusion **N**etwork, for region-specific traffic flow. In ST-GDN, we develop a multi-scale self-attention network to investigate multi-

*Corresponding author: Chao Huang

grained temporal dynamics across various time resolutions, in order to encode temporal hierarchy of traffic transitional regularities. To promote the collaboration of different resolution-aware temporal representations, an aggregation layer is proposed to model the underlying dependencies across multi-level temporal dynamics. In addition, the developed hierarchical graph neural network via attentive graph diffusion paradigm, endows the ST-GDN with the capability to incorporate spatial semantics from local-level spatially adjacent relations to global-level traffic pattern representations across the city in a joint manner.

We highlight the key contributions of this work as:

- We highlight the critical importance of explicitly exploring the multi-resolution traffic transitional information and local-global cross-region dependencies, in studying the traffic prediction problem.
- We propose a new traffic prediction framework (ST-GDN) which explicitly embeds multi-level temporal contextual signals into resolution-aware latent representations, with the cooperation of the designed multi-scale self-attention network and temporal hierarchy aggregation layer.
- ST-GDN preserving both local and global region-wise dependencies, via a hierarchically structured graph neural architecture which is integrated with a graph attention network and convolution-based graph diffusion mechanism.
- Our extensive experiments on four real-world datasets demonstrate that ST-GDN outperforms baselines of different types in yielding better forecasting performance. Furthermore, model efficiency study is conducted for ST-GDN and several compared methods.

## Problem Definition

In this section, we begin with some key definitions and preliminary terms. Then, we present our studied task of traffic flow forecasting.

**Definition 1** ***Spatial Region***. *We partition a city into $I \times J$ disjoint grids (given the geographical coordinates), in which each grid is regarded as a spatial region $r_{i,j}$ ($i \in [1, ..., I]$, $j \in [1, ..., J]$). $r_{i,j}$ is our target unit for traffic prediction.*

**Definition 2** ***Traffic Flow Tensor***. *After the grid-based partition, we represent the citywide traffic volume distributions across regions during past $T$ time slots as a three-way tensor: $\boldsymbol{X} \in \mathbb{R}^{I \times J \times T}$, where each entry $x_{i,j}^t$ denotes the traffic volume measurement at region $r_{i,j}$ in the $t$-th time slot (e.g., hour or day). To study the prediction on both the incoming and outgoing traffic follow, we generate two traffic flow tensors: $\boldsymbol{X}^\alpha$ (incoming) and $\boldsymbol{X}^\beta$ (outgoing), respectively.*

**Task Formulation**. Based on the aforementioned definitions, the traffic prediction problem is formulated as: **Input**: the observed traffic volume information during past $T$ time slots across the entire city $\mathbf{X}^\alpha \in \mathbb{R}^{I \times J \times T}$ and $\mathbf{X}^\beta \in \mathbb{R}^{I \times J \times T}$. **Output**: a predictive function which effectively infers the unknown traffic volume in future time slots.

## Methodology

In this section, we elaborate our proposed ST-GDN framework with the technical details (as shown in Figure 1).

### Temporal Hierarchy Modeling

We first propose a multi-scale self-attention network to jointly map multi-level temporal signals into common latent representations, for capturing the complex traffic patterns.

**Definition 3** ***Temporal Resolution*** *$p$. We define $p$ to indicate how often we sample traffic volume measurement $x_{i,j}^t$ from the overall traffic flow tensor $\boldsymbol{X}$, i.e., the time difference between two consecutive data points $x_{i,j}^t$ and $x_{i,j}^{t'}$ measured from region $r_{m,n}$. For example, $(t' - t)$ can be a hour, a day or a week, given the resolution $p$ is set as hourly, daily and weekly, respectively, i.e., $p \in \{hour, day, week\}$.*

Given each temporal resolution $p$, we could generate resolution-aware traffic series $\mathbf{x}_{i,j}^{T_p}$, where $T_p$ is the corresponding traffic series length with the resolution of $p$. Then, we propose a self-attentive network to encode the traffic variation patterns from the temporal dimension. In particular, our encoder is built upon the scaled dot-product attention architecture with three transformation matrices: query ($\mathbf{Q} \in \mathbb{R}^{T_p \times d}$), key ($\mathbf{K} \in \mathbb{R}^{T_p \times d}$) and value ($\mathbf{V} \in \mathbb{R}^{T_p \times d}$) matrices. The resolution-aware attentive aggregation mechanism can be formally presented with the matrix calculation:

$$\begin{bmatrix} \mathbf{Q} \\ \mathbf{K} \\ \mathbf{V} \end{bmatrix} = \mathbf{E}^p \begin{bmatrix} \mathbf{W}_Q \\ \mathbf{W}_K \\ \mathbf{W}_V \end{bmatrix}; \mathbf{Y}^p = \sigma(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V} \tag{1}$$

where $\mathbf{y}_{i,j}^p \in \mathbf{Y}^p$ and $\mathbf{y}_{i,j}^p \in \mathbb{R}^d$ denotes the learned resolution-aware hidden representation of region $r_{i,j}$. $E_p \in \mathbb{R}^{|R| \times d}$ is the initialized embeddings of all regions $r_{i,j} \in R$. Additionally, $\sigma(\cdot)$ denotes the softmax function.

### Traffic Dependency Learning with Global Context

The goal of this step is to exploit the global-level dependencies across different regions in terms of their dynamic traffic transition patterns. Towards this end, we first define a region graph $G = (R, E)$, in which $R$ is the region set and $E$ denotes the pairwise relationships between two spatial regions. Motivated by the attention neural network in encoding the dependencies among regions (Huang et al. 2019), we develop an attentive aggregation mechanism to capture both local and global traffic dependency between regions. Specifically, we perform the message aggregation over $G$ with the following attentive operations.

$$m_{(i,j) \leftarrow (i',j')}^p = \Big\Vert_{h=1}^{H} \omega_{(i,j);(i',j')}^h \cdot \mathbf{Y}^p \cdot \mathbf{W}^p \tag{2}$$

where $m_{(i,j) \leftarrow (i',j')}^p$ is the feature message propagated from region $r_{i',j'}$ to $r_{i,j}$. Here, we endow the cross-region relevance encoding with multi-head ($h \in [1, ..., H]$), to capture the region-wise relation semantic from different learning subspaces. Furthermore, $\mathbf{W}^p \in \mathbb{R}^{d \times d}$ is the parameterized projection matrix. the underlying attentive relevance

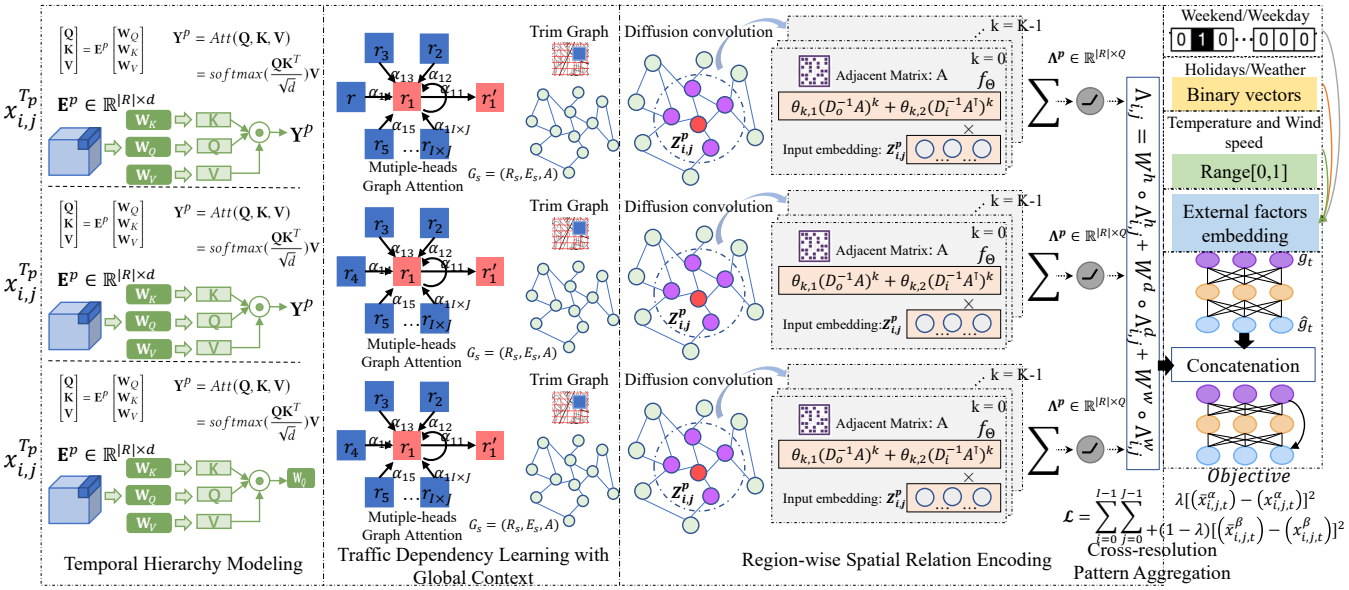



Figure 1: The framework of our developed spatial-temporal graph diffusion networks.

$\omega^h_{(i,j);(i',j')}$ is formally estimated as follows:

$$\omega^h_{(i,j);(i',j')} = \frac{exp(LR(\boldsymbol{\alpha}^T[\widetilde{\mathbf{y}}^p_{i,j}||\widetilde{\mathbf{y}}^p_{i',j'}]))}{\sum_{(i',j')\in\mathcal{N}(i,j)} exp(LR(\boldsymbol{\alpha}^T[\widetilde{\mathbf{y}}^p_{i,j}||\widetilde{\mathbf{y}}^p_{i',j'}]))}$$

where we perform concatenation between $\widetilde{\mathbf{y}}^p_{i',j'}$ and $\widetilde{\mathbf{y}}^p_{i',j'}$ ($\widetilde{\mathbf{y}}^p_{i',j'} = \mathbf{y}^p_{i',j'} \cdot \mathbf{W}^p$). Then, the attentive coefficient vector $\alpha$ is incorporated with the production. $LR(\cdot)$ denotes the *LeakyReLU* function. Based on the constructed message and learned quantitative region-wise relevance score $\omega_{(i,j);(i',j')}$, we perform the information aggregation as:

$$\mathbf{z}^p_{i,j} = f\Big(\sum_{r_{i',j'}\in\mathcal{N}_{i,j}} m^p_{(i,j)\leftarrow(i',j')}\Big) \tag{3}$$

where $\mathbf{z}^p_{i,j}$ is the aggregated feature embedding of $r_{i,j}$.

**High-order Information Propagation**. The information aggregation from the $(l)$-th layer to the $(l+1)$-th layer with the high-order relation modeling is represented as:

$$\mathbf{z}^{p,(l+1)}_{i,j} \leftarrow \underset{i\in N_u(j);j'\in N_v(j)}{\textbf{Aggregate}} \Big(\textbf{Propagate}(\mathbf{z}^{p,(l)}_{i,j}, G)\Big) \tag{4}$$

Propagate$(\cdot)$ and Aggregate$(\cdot)$ denotes the message construction and information fusion, respectively. We finally generate the global-level representation of region $r_{i,j}$ as: $\mathbf{z}^p_{i,j} = \mathbf{z}^{p,(l)}_{i,j} \oplus ... \oplus \mathbf{z}^{p,(L)}_{i,j}$. $\oplus$ is the element-wise addition.

## Region-wise Relation Learning with Graph Diffusion Paradigm

In addition to the global dependencies across different regions in terms of their traffic evolving patterns, we further incorporate spatial relationships between regions into our prediction framework. Particularly, we develop a graph-structured diffusion network to refine the learned resolution-aware region representations $\mathbf{z}^p_{i,j}$ from the above graph attention module. We generate another region-wise relation graph $G_s = (R_s, E_s, A)$ which jointly preserves the geographical adjacent relations ($r_{i,j}$'s $\sqrt{K} \times \sqrt{K} = K$ neighboring regions) and high traffic dependencies (larger $\omega_{(i,j);(i',j')}$ value). $A$ denotes the adjacent matrix which is weighted by a vertex distance function. Here, we define $D_o = \mathbf{A} \cdot \mathbf{I}$ to denote the out-degree diagonal matrix, where $\mathbf{I}$ is the identify matrix of $G_s$. The designed diffusion convolution operation performs the diffusion process across each vertex in graph $G_s$ to generate new feature representations as:

$$f(\mathbf{z}^p_{i,j})_\Theta = \sum_{k=0}^{K-1} (\theta_{k,1}(D_o^{-1}\mathbf{A})^k + \theta_{k,2}(D_i^{-1}\mathbf{A}^\intercal)^k)\mathbf{z}^p_{i,j} \tag{5}$$

where $\theta_{k,1}, \theta_{k,2} \in \mathbb{R}^{K\times 2}$. $D_o^{-1}\mathbf{A}$ (in-degree) and $D_i^{-1}\mathbf{A}^\intercal$) (out-degree) denote the bi-directional transition matrices of the diffusion process, which corresponds to the inflow and outflow in our prediction scenario. The parameter tensor denoted as $\Theta \in \mathbb{R}^{Q\times d\times K\times 2}$, in which the Q-dimensional output $\mathbf{\Lambda}^p \in \mathbb{R}^{|R|\times Q}$ of diffusion convolutional layer is given:

$$\mathbf{\Lambda}^p_q = LeakyReLU\Big(\sum_{d'=1}^{d} f(\mathbf{Z}^p_{d'})_{\Theta_{q,d'}}\Big) \tag{6}$$

where $q \in \{1, ..., Q\}$. The obtained region representation $\mathbf{\Lambda}^p_{i,j}$ jointly preserves the temporal (traffic time-varying patterns) and spatial (geographical relations) contextual signals under a global perspective.

We next aggregate the resolution-aware traffic representation $\mathbf{\Lambda}^p_{i,j}$ by introducing a gating mechanism. To be specific, our gated aggregation mechanism conducts the parametric matrix-based sum operation over the multi-resolution traffic pattern representations, *i.e.*, hourly ($\mathbf{\Lambda}^{p_h}$), daily ($\mathbf{\Lambda}^{p_d}$) and weekly ($\mathbf{\Lambda}^{p_w}$) as follows:

$$\mathbf{\Lambda}_{i,j} = \mathbf{W}^h \circ \mathbf{\Lambda}^h_{i,j} + \mathbf{W}^d \circ \mathbf{\Lambda}^d_{i,j} + \mathbf{W}^w \circ \mathbf{\Lambda}^w_{i,j} \tag{7}$$

Here, the trainable transformation matrices are denoted as $\mathbf{W}^h$, $\mathbf{W}^d$ and $\mathbf{W}^w$ corresponding to hourly, daily and weekly

patterns. We finally generate the conclusive multi-resolution traffic representation $\mathbf{\Lambda}_{i,j}$ which preserves multi-grained temporal hierarchy of traffic regularities.

### Traffic Prediction Phase

In the urban sensing scenario, there exist external factors (*e.g.*, meteorological conditions) which impact traffic transitional regularities. Thus, we further augment our ST-GDN with the capability of fusing heterogeneous external factors. In particular, we consider three categories of external factors: Weather conditions, Temperature/°C, Wind speed/mph. We follow the similar strategies in (Liang et al. 2018) for mapping these features into vectors $\mathbf{g}_t$. After that, we utilize a multi-layer perceptron framework to perform projection over $\hat{\mathbf{g}}_t$. Finally, we feed the concatenated embedding ($\mathbf{\Lambda}_{i,j}$ and $\hat{\mathbf{g}}_t$) into the prediction layer to infer the traffic volume.

**Optimized Loss Function**. We define our loss function with the joint consideration of inflow and outflow traffic volume of each region in a city as below:

$$\begin{aligned}\mathcal{L} = \sum_{i=0}^{I-1}\sum_{j=0}^{J-1} \lambda[(\bar{x}^{\alpha}_{i,j,t}) - (x^{\alpha}_{i,j,t})]^2 \\ +(1-\lambda)[(\bar{x}^{\beta}_{i,j,t}) - (x^{\beta}_{i,j,t})]^2\end{aligned} \tag{8}$$

where $\bar{x}^{\alpha}_{i,j,t}$ and $\bar{x}^{\beta}_{i,j,t}$ denotes the estimated incoming and outgoing traffic volume of region $r_{i,j}$ at the $t$-th time slot, respectively. Their influences are decided by $\lambda$. Ground truth information are represented $x^{\alpha}_{i,j,t}$ and $x^{\beta}_{i,j,t}$.

**Model Complexity Analysis**. We analyze the time complexity of our ST-GDN framework. Particularly, the multi-scale self-attentive network takes $O(3 \times T \times I \times J \times d)$ for learning query, key and value matrices, and $O(3 \times T^2 \times d)$ for weighted summation. The next graph attention module takes $O(3 \times I^2 \times J^2 \times d')$ to estimate the relevance scores and perform feature aggregation, which dominates the computational cost of our ST-GDN. Additionally, the graph diffusion-based spatial relation modeling takes $O(K \times |E_s|)$. The external factor fusion takes the $O(3 \times d)$ complexity.

## Evaluation

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

- **RQ1**: How is the overall traffic prediction performance of *ST-GDN* as compared to various baselines?
- **RQ2**: How do designed different sub-modules contribute to the model performance?
- **RQ3**: How does *ST-GDN* perform *w.r.t* different time granularity configurations for temporal context modeling?
- **RQ4**: What is the influence of hyperparameter settings?
- **RQ5**: How is the model efficiency of *ST-GDN*?

### Experimental Settings

**Data Description.** Our experiments are performed on four real-world traffic datasets, which are summarized in Table 1:

**BJ-Taxi** (Zhang, Zheng, and Qi 2017). There are 34,000+ processed taxi trajectories included in this data. Each trajectory is mapped into one of $32 \times 32$ grid-based geographical regions. The traffic volume is measured every half an hour.

**NYC-Taxi** (Yao et al. 2019). This data contains 22,000,000+ taxi trajectories collected from 01/01/2015 to 03/01/2015 in New York City with a $10 \times 20$ grid map. The traffic data sample period is also half an hour.

**NYC-Bike-1** (Zhang, Zheng, and Qi 2017). It includes the trajectories of the bike system from New York with a $16 \times 8$ grid map. Traffic volume is estimated on a hourly basis.

**NYC-Bike-2** (Yao et al. 2019). It is another bike trajectory data which spans from 07/01/2016 to 08/29/2016 in NYC. The whole data consists of 2,600,000+ trajectory records which are collected with $10 \times 20$ grid map. The data measurement interval is 30 mins.

**Evaluation Protocols.** In our experiments, we leverage two representative metrics for evaluation: *Root Mean Squared Error (RMSE)* and *Mean Absolute Percentage Error (MAPE)* (Liang et al. 2019). We present the partition details of training/validation/test datasets in Table 2. Validation set gives an estimate of model skill while tuning model's hyperparameters with the data held back from training set.

**Methods for Comparison.** In the performance comparison between our method and state-of-the-art traffic forecasting techniques, we consider the following baselines with various model structures.

**Traditional Time Series Prediction Approaches**:

- **ARIMA** (Pan, Demiryurek, and others 2012). it is a representative method for forecasting time series data.
- **Support Vector Regression (SVR)** (Chang and Lin 2011): another traditional time series analysis model via learning feature mapping functions.

**Conventional Hybrid Learning Approach**:

- **Fuzzy+NN** (Srinivasan, Chan, and Balaji 2009): it integrates the feed-forward neural layers with the fuzzy input filter to model the traffic patterns.

**Recurrent Spatial-Temporal Prediction Methods**:

- **ST-RNN** (Liu et al. 2016): it leverages the recurrent neural networks for capturing both the spatial and temporal effects for making sequential data prediction.
- **D-LSTM** (Yu et al. 2017a): it jointly models the normal and abnormal traffic variations based on stacked long short-term memory networks.

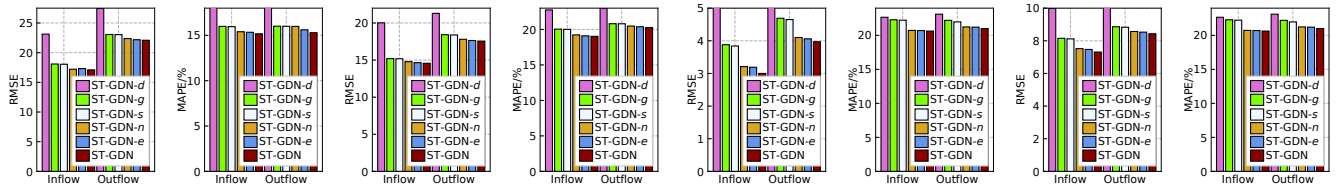
**Convolution-based Network for Traffic Forecasting**:

- **DeepST** (Zhang et al. 2016): it utilizes the convolution neural network to encode the spatial correlations between regions over a citywide grid map.
- **ST-ResNet** (Zhang, Zheng, and Qi 2017): the residual connection technique is employed to alleviate overfitting issue for spatial-temporal prediction.

**Convolutional Recurrent Predictive Solution**:

Table 1: Statistical information of experimented datasets.

| Dataset | BJ-Taxi | NYC-Taxi | NYC-Bike-1 | NYC-Bike-2 | External Factors | Beijing | New York City |
|---|---|---|---|---|---|---|---|
| Data type | Taxi GPS | Taxi GPS | Bike Rental | Bike Rental | Weather Condition | sunny, rainy, foggy, snowy | sunny, rainy, foggy, snowy |
| Time interval | 30 minutes | 30 minutes | one hour | 30 minutes | Temperature/°C | $[-24.6, 41.0]$ | $[-10.3, 31.40]$ |
| Gird map size | 32×32 | 10×20 | 16×8 | 10×20 | Wind speed/mph | $[0, 48.60]$ | $[0, 63.75]$ |
| Number of records | 34,000+ | 22,000,000+ | 6,800+ | 2,600,000+ | Holidays | weekends, national Holidays | weekends, national Holidays |




(a) NYC-Taxi (b) NYC-Taxi (c) BJ-Taxi (d) BJ-Taxi (e) NYC-Bike-1 (f) NYC-Bike-1 (g) NYC-Bike-2 (h) NYC-Bike-2

Figure 2: Model ablation study of ST-GDN framework in terms of RMSE and MAPE.

Table 2: Training/validation/test data split details.

| Data | Training | Validation | Test |
|---|---|---|---|
| BJ-Taxi | 7/1/2013-12/31/2015 | 1/1/2016-1/21/2016 | 1/22/2016-4/10/2016 |
| NYC-Taxi | 1/1/2015-2/14/2015 | 2/15/2015-2/19/2015 | 2/20/2015-3/1/2015 |
| NYC-Bike-1 | 4/1/2014-9/15/2014 | 9/16/2014-9/20/2014 | 9/21/2014-9/30/2014 |
| NYC-Bike-2 | 7/1/2016-8/14/2016 | 8/15/2016-8/19/2016 | 8/20/2016-8/29/2016 |

- **DMVST-Net** (Yao et al. 2018): it integrates the graph embedding method with the joint convolutional recurrent networks to capture spatial-temporal signals
- **DCRNN** (Li et al. 2018): it is a data-driven forecasting framework with diffusion recurrent neural network to capture the spatial-temporal dependencies.

**Attentive Traffic Prediction Model**:

- **STDN** (Yao et al. 2019): it designs a periodically shifted attention for learning transition regularities of traffic.

**Traffic Prediction with Graph Neural Networks**:

- **ST-GCN** (Yu, Yin, and Zhu 2018): it is an integrative framework of graph convolution network and convolutional sequence modeling layer for modeling spatial and temporal dependencies.
- **ST-MGCN** (Geng et al. 2019): it develops a multi-modal graph convolutional network to capture region-wise non-Euclidean pair-wise correlations.
- **GMAN** (Zheng et al. 2020): it is a encoder-decoder traffic prediction method based on the graph multi-attention.

**Deep Hybrid Traffic Flow Predictive Models**:

- **UrbanFM** (Liang et al. 2019): it is a deep fusion network to model traffic flow distributions.
- **ST-MetaNet** (Pan et al. 2019): it is a meta-learning approach to perform knowledge transfer across series with a recurrent graph attentive network.

**Parameter Settings.** The *ST-GDN* is implemented with Tensorflow. The training phase is performed using the Adam optimizer with the learning rate of $1e^{-3}$ and batch size of 32. The embedding dimension size $d$ and the depth recursive graph neural layers $L$ are set as 64 and 3, respectively. We select the input sequence length from the range of $\{1, 2, 3, 4, 5, 6\}$, $\{1, 2, 3, 4, 5\}$, $\{1, 2, 3, 4, 5, 6\}$, which respectively corresponds to three different time resolutions (hour–$T_h$, day–$T_d$ and week–$T_w$). We stack three feed-forward layers in the final prediction phase. The experiments of most baselines are performed with their released code.

## Performance Comparison (RQ1)

**Performance Superiority of *ST-GDN***. The performance comparison results of all methods are presented in Table 3. We can observe that *ST-GDN* consistently yields the best performance in all cases, which demonstrates the effectiveness of our *ST-GDN* in jointly modeling of multi-level temporal dynamics and global-level region-wise dependencies. Figure 3 visualize the prediction error ($[(\bar{x}_{i,j,t}) - (x_{i,j,t})]^2$) of our *ST-GDN* and five best performed baselines on BJ-taxi data, where a brighter pixel means a larger error. The superiority of *ST-GDN* can still be observed, which is consistent with the quantitative results in Table 3.

**Performance Comparison between Baselines**. Compared with conventional time series approaches, neural network-based models perform better in most evaluation cases. The subsequent attention-based and recurrent-convolutional network methods (*e.g.*, STDN, DMVST-Net) obtain better performance than recurrent neural models (*e.g.*, D-LSTM), which justifies the necessity to simultaneously capture both spatial and temporal relations in traffic prediction. Among various baselines, GNN-based methods have better performance than other types of competitors, which ascertains the rationality of designing graph-structured information aggregation mechanism to fuse spatial and temporal signals.

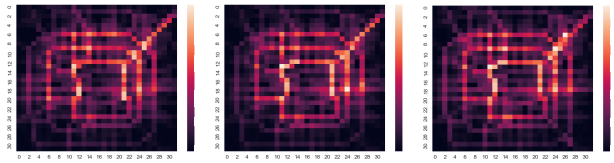
## Comparison with Variants (RQ2)

We perform ablation experiments to analyze the effects of sub-modules in our *ST-GDN* framework with five variants:

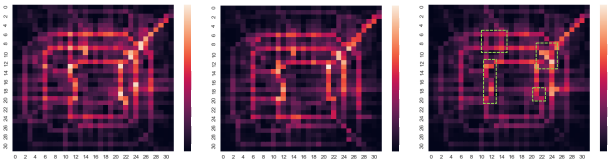
- *ST-GDN*-s: *ST-GDN* without the multi-scale self-attention network to capture multi-level traffic dynamics.
- *ST-GDN*-g: *ST-GDN* without the graph attention module to model the global region-wise traffic dependencies.
- *ST-GDN*-d: *ST-GDN* without the graph diffusion network to integrate spatial context with cross-region traffic pattern correlations for representation recalibration.

Table 3: Performance comparison of all methods on four datasets in terms of *RMSE* and *MAPE*.

| Datasets | BJ-Taxi | | | | NYC-Bike-1 | | | | NYC-Taxi | | | | NYC-Bike-2 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Metrics | RMSE | | MAPE (%) | | RMSE | | MAPE (%) | | RMSE | | MAPE (%) | | RMSE | | MAPE (%) | |
| Methods | In | Out | In | Out | In | Out | In | Out | In | Out | In | Out | In | Out | In | Out |
| ARIMA | 22.10 | 24.01 | 30.89 | 32.24 | 9.30 | 11.81 | 35.82 | 36.47 | 27.21 | 36.54 | 20.90 | 22.18 | 11.26 | 11.52 | 25.74 | 26.56 |
| SVR | 21.44 | 22.12 | 22.64 | 22.32 | 8.65 | 9.07 | 23.58 | 24.10 | 26.16 | 34.71 | 18.25 | 21.01 | 10.10 | 11.03 | 23.47 | 24.01 |
| Fuzzy+NN | 22.35 | 23.06 | 22.67 | 22.73 | 8.56 | 9.17 | 24.03 | 24.48 | 25.98 | 34.50 | 18.92 | 21.54 | 11.31 | 11.84 | 24.70 | 25.15 |
| ST-RNN | 27.16 | 27.90 | 24.17 | 24.72 | 8.99 | 9.24 | 28.22 | 28.58 | 29.88 | 37.23 | 25.97 | 26.55 | 13.51 | 15.39 | 27.06 | 27.60 |
| D-LSTM | 26.99 | 27.56 | 23.64 | 24.17 | 8.64 | 9.10 | 27.51 | 28.07 | 29.52 | 37.04 | 25.81 | 26.11 | 13.64 | 15.89 | 27.21 | 27.88 |
| DeepST | 19.30 | 21.06 | 22.45 | 22.52 | 7.66 | 8.16 | 22.81 | 23.21 | 23.56 | 26.79 | 22.34 | 22.39 | 7.60 | 8.15 | 22.78 | 23.18 |
| ST-ResNet | 17.00 | 22.31 | 23.51 | 23.74 | 6.28 | 6.61 | 23.92 | 24.79 | 21.72 | 26.30 | 21.12 | 21.24 | 8.84 | 9.85 | 23.05 | 23.15 |
| DMVST-Net | 16.61 | 17.14 | 22.52 | 23.06 | 5.82 | 6.09 | 22.45 | 23.67 | 20.63 | 25.80 | 17.19 | 17.44 | 8.70 | 9.31 | 21.72 | 22.35 |
| STDN | 15.19 | 18.63 | 21.04 | 22.13 | 4.50 | 5.92 | 21.71 | 22.61 | 19.31 | 24.19 | 16.43 | 16.59 | 8.25 | 9.00 | 21.23 | 22.24 |
| UrbanFM | 15.18 | 18.42 | 20.54 | 20.88 | 3.99 | 4.64 | 21.59 | 22.47 | 19.11 | 24.14 | 16.34 | 16.46 | 8.19 | 8.88 | 21.25 | 22.22 |
| ST-MetaNet | 15.06 | 18.29 | 19.91 | 20.74 | 3.85 | 4.64 | 21.26 | 22.18 | 18.30 | 23.88 | 16.19 | 16.27 | 8.13 | 8.82 | 21.18 | 21.72 |
| DCRNN | 15.13 | 18.37 | 20.14 | 20.88 | 3.86 | 4.65 | 21.14 | 21.05 | 18.19 | 23.74 | 16.11 | 16.16 | 8.15 | 8.83 | 21.21 | 21.94 |
| ST-GCN | 15.11 | 18.30 | 19.92 | 20.77 | 3.76 | 4.70 | 21.12 | 21.94 | 18.02 | 23.08 | 15.94 | 15.92 | 8.00 | 8.74 | 21.18 | 21.91 |
| ST-MGCN | 15.08 | 18.25 | 19.96 | 20.70 | 3.75 | 4.63 | 21.04 | 21.95 | 17.97 | 23.00 | 15.87 | 15.91 | 7.92 | 8.72 | 21.20 | 21.71 |
| GMAN | 15.07 | 18.23 | 19.97 | 20.68 | 3.73 | 4.64 | 21.02 | 21.93 | 17.95 | 22.96 | 15.84 | 15.89 | 7.88 | 8.73 | 21.18 | 21.70 |
| ***ST-GDN*** | **14.57** | **17.56** | **19.03** | **20.27** | **3.00** | **3.97** | **20.48** | **21.31** | **17.10** | **22.09** | **15.17** | **15.29** | **7.31** | **8.43** | **20.63** | **21.00** |



(a) ST-GCN (b) ST-MGCN (c) ST-MetaNet



(d) DCRNN (e) GMAN (f) ST-GDN

Figure 3: Visualization for Traffic Prediction Errors.

- *ST-GDN*-n: *ST-GDN* without the incorporation of neighborhood spatial context into the graph diffusion.
- *ST-GDN*-e: *ST-GDN* without the external factor fusion.

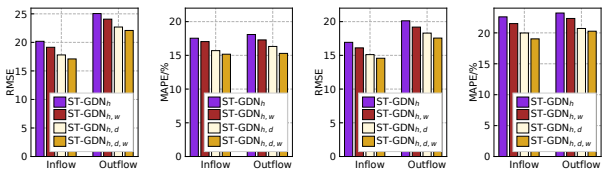
The evaluation results are shown in Figure 2. We can observe that the joint version of *ST-GDN* outperforms other variants consistently. Hence, each designed sub-modules has positive effects for prediction performance improvement. It is necessary to build a joint framework to collectively integrate the multi-resolution traffic temporal patterns, global region-wise traffic dependencies, and regions' geographical relations, into the spatial-temporal traffic pattern modeling.

## Multi-Resolution Temporal Effects (RQ3)

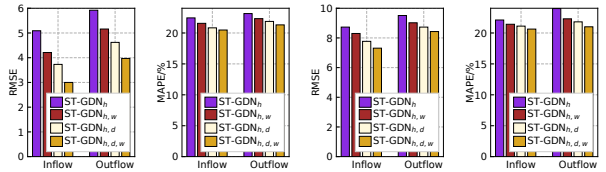
In this subsection, we study the effects of different temporal resolution settings in our integrative architecture of multi-scale self-attention network and cross-resolution pattern aggregation layer, with the following contrast models:

- *ST-GDN*$_h$: $P \in \{hour/30mins\}$
- *ST-GDN*$_{h,d}$: $P \in \{hour/30mins, day\}$
- *ST-GDN*$_{h,w}$: $P \in \{hour/30mins, week\}$
- *ST-GDN*$_{h,d,w}$: $P \in \{hour/30mins, day, week\}$

We present the study results in Figure 4. As we can seen, the best prediction accuracy is achieved by *ST-GDN*$_{h,d,w}$ which is configured with more resolutions. Leaning the temporal hierarchy with hourly and daily/weekly traffic patterns (*ST-GDN*$_{h,d}$, *ST-GDN*$_{h,w}$) provide better results as compared to the variant with singular-dimensional time granularity (*ST-GDN*$_h$). Overall, decomposing the temporal effects into more multiple resolution-specific feature representations is helpful for more accurate modeling of traffic temporal regularity and resolution-aware region relations.




(a) NYC-Taxi (b) NYC-Taxi (c) BJ-Taxi (d) BJ-Taxi




(e) NYC-Bike-1 (f) NYC-Bike-1 (g) NYC-Bike-2 (h) NYC-Bike-2

Figure 4: Multi-resolution temporal effect studies.

## Parameter Sensitivity (RQ4)

**Depth of Graph Attention Network** $L$. We can notice that increasing the depth of our graph attention module by stacking multiple embedding propagation layers could boost the performance. The results also indicate that exploring third-order relations among region entities is sufficient to capture the global traffic dependencies.

**Length of Encoded Input Sequence** $T$. The performance is initially improved with the increase of $T_h$ and $T_d$, since longer traffic series can provide more useful temporal information. However, the further increasing of sequence length may introduce noise which mislead the traffic modeling.

**Kernel Size** $K$. We vary the kernel size to investigate the convolution operations in our graph diffusion process. We can observe that $K = 3$ achieves the best performance.

**# of Sampled Neighbor Regions**. As we increase the size of neighbor sample grid map to $3 \times 3$, a larger geographical coverage results in better performance. However, the performance degrades with $4 \times 4$ and $5 \times 5$. The reason is that the training of *ST-GDN* becomes harder with more parameters are involved when modeling more neighboring relations.

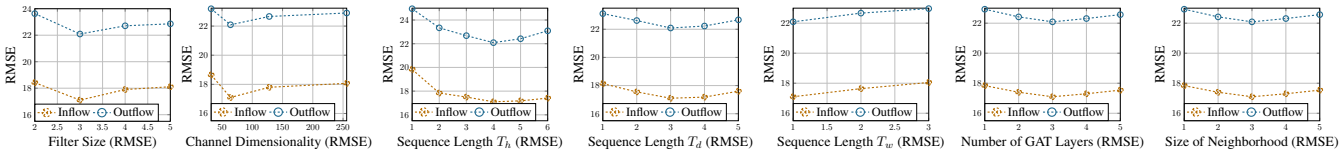



Figure 5: Hyper-parameter study on NYC-Taxi data in terms of RMSE.

Table 4: Model Efficiency Study.

| Methods | Training | | | |
|---|---|---|---|---|
| | BJ-Taxi | NYC-Taxi | NYC-Bike-1 | NYC-Bike-2 |
| ST-MetaNet | 16121.01 | 1298.55 | 1020.14 | 1218.33 |
| DCRNN | 7996.24 | 981.36 | 705.64 | 938.17 |
| ST-GCN | 4088.90 | 744.65 | 500.40 | 732.11 |
| ST-MGCN | 8263.27 | 1023.29 | 789.71 | 1006.46 |
| GMAN | 7368.31 | 854.66 | 547.12 | 781.66 |
| *ST-GDN* | 7625.19 | 891.63 | 569.26 | 791.52 |
| Methods | Inference | | | |
| | BJ-Taxi | NYC-Taxi | NYC-Bike-1 | NYC-Bike-2 |
| ST-MetaNet | 0.42 | 0.32 | 0.29 | 0.30 |
| DCRNN | 0.26 | 0.24 | 0.21 | 0.22 |
| ST-GCN | 0.25 | 0.22 | 0.19 | 0.21 |
| ST-MGCN | 0.31 | 0.27 | 0.25 | 0.26 |
| GMAN | 0.25 | 0.23 | 0.19 | 0.22 |
| *ST-GDN* | 0.26 | 0.23 | 0.20 | 0.22 |

**Channel Dimensionality**. The results suggest that larger channel dimension size does not always bring the stronger representation ability, due to the overfitting issue.

## Model Efficiency Study (RQ5)

We finally investigate the model efficiency (measured by running time) of our *ST-GDN*. All experiments are conducted with the default parameter configurations on a single NVIDIA GeForce GTX 1080 Ti GPU. We observe that in several best performed baselines, *ST-GCN* has good prediction accuracy and running speed. Our *ST-GDN* outperforms most of compared approaches and could achieve competitive efficiency as compared to *ST-GCN*, *i.e.*, the attention-based graph embedding propagation layer has higher computational cost than the adjacent matrix-based graph convolution. Considering the prediction accuracy comparison between *ST-GDN* and *ST-GCN*, the additional computational cost could bring positive effect via learning global region inter-dependencies in an explicit manner.

We finally investigate the model efficiency (measured by running time) of our *ST-GDN*. Table 4 presents the computational cost of training (with 300 epochs) and inference phase for *ST-GDN* and five best performed baselines on four different datasets. All experiments are conducted with the default parameter configurations on a single NVIDIA GeForce GTX 1080 Ti GPU. We can observe that *ST-GDN* outperforms most of compared approaches and could achieve competitive efficiency as compared to *ST-GCN*, *i.e.*, the attention-based graph embedding propagation layer has higher computational cost than the adjacent matrix-based graph convolution. Considering the prediction accuracy comparison between *ST-GDN* and *ST-GCN*, the additional computational cost could bring positive effect via learning global region inter-dependencies in an explicit manner.

## Related Work

**Traffic Prediction with Deep Learning**. Recently, many efforts have been devoted to developing traffic prediction techniques based on various neural network architectures. One straightforward solution is to apply the recurrent neural networks (*e.g.*, LSTM) to encode the temporal features of traffic series (Yu et al. 2017a; Liu et al. 2016). The subsequent extensions propose to integrate the recurrent neural layers with the convolutional network (Zhang, Zheng, and Qi 2017; Yao et al. 2018) or attention mechanism (Yao et al. 2019), so as to joint model the spatial-temporal signals. In addition, some hybrid methods have been proposed for traffic prediction with the exploration of heterogeneous data fusion (Liang et al. 2019) and meta-learning-based knowledge transfer (Pan et al. 2019). Different from these work, ST-GDN endows the spatial-temporal pattern representation process with the preservation of hierarchical temporal dynamics and global-enhanced region-wise dependencies. While there exist research work that considers the global dependency among regions (Zhang et al. 2020), it is limited in its separately modeling of traffic dependency and nearby region relations. In contrast, ST-GDN incorporates the global context enhanced region-wise explicit relevance into a graph diffusion paradigm to capture comprehensive inter-region dependencies in a joint learning manner.

**Graph Neural Networks**. It is worth mentioning that several recent efforts have investigated GNNs for spatial-temporal data forecasting (Guo, Lin, and others 2019; Song et al. 2020). For example, ST-GCN (Yu, Yin, and Zhu 2018) and ST-MGCN (Geng et al. 2019) proposes to leverage graph convolution network to model correlations between regions. Furthermore, attention mechanism has been introduced for information aggregation from adjacent roads (Zheng et al. 2020; Wang et al. 2020). Motivated by these work, we develop a hierarchical graph neural architectures to promote the cooperation between the multi-resolution temporal context with the dual-modal cross-region inter-dependencies, which have not been well explored in existing solutions.

## Conclusion

This work investigates the traffic prediction problem by proposing a new architecture (ST-GDN) based graph neural networks. Specifically, it first designs a resolution-aware self-attention network to encode the multi-level temporal signals. Then, the local spatial contextual information and global traffic dependencies across different regions, are subsequently integrated to enhance the spatial-temporal pattern representations. Comprehensive experiments demonstrate that the proposed ST-GDN significantly outperforms 15 baselines over four datasets consistently. Our future work lies in the deployment of our developed prototype in a cloud-based working system for real-time traffic flow prediction.

## Acknowledgments

The authors would like to thank the anonymous referees for their valuable comments and helpful suggestions.

This work is supported by National Nature Science Foundation of China (62072188, 61672241), Natural Science Foundation of Guangdong Province (2016A030308013), Science and Technology Program of Guangdong Province (2019A050510010). This work is also partially supported by National Key RD Program of China (2019YFB2101801) and the Beijing Nova Program (Z201100006820053).

## References

[Chang and Lin 2011] Chang, C.-C., and Lin, C.-J. 2011. Libsvm: a library for support vector machines. *TIST* 2(3):27.

[Deng et al. 2016] Deng, D.; Shahabi, C.; Demiryurek, U.; Zhu, L.; et al. 2016. Latent space model for road networks to predict time-varying traffic. In *KDD*, 1525–1534.

[Diao et al. 2019] Diao, Z.; Wang, X.; Zhang, D.; Liu, Y.; Xie, K.; and He, S. 2019. Dynamic spatial-temporal graph convolutional neural networks for traffic forecasting. In *AAAI*, volume 33, 890–897.

[Gao et al. 2019] Gao, Y.; Zhao, L.; Wu, L.; Ye, Y.; Xiong, H.; and Yang, C. 2019. Incomplete label multi-task deep learning for spatio-temporal event subtype forecasting. In *AAAI*, volume 33, 3638–3646.

[Geng et al. 2019] Geng, X.; Li, Y.; Wang, L.; Zhang, L.; Yang, Q.; et al. 2019. Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting. In *AAAI*.

[Guo, Lin, and others 2019] Guo, S.; Lin, Y.; et al. 2019. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *AAAI*, volume 33, 922–929.

[Huang et al. 2018] Huang, C.; Zhang, J.; Zheng, Y.; and Chawla, N. V. 2018. Deepcrime: attentive hierarchical recurrent networks for crime prediction. In *CIKM*, 1423–1432.

[Huang et al. 2019] Huang, C.; Zhang, C.; Zhao, J.; Wu, X.; Yin, D.; and Chawla, N. 2019. Mist: A multiview and multimodal spatial-temporal learning framework for citywide abnormal event forecasting. In *WWW*, 717–728.

[Huang et al. 2020] Huang, C.; Zhang, C.; Dai, P.; and Bo, L. 2020. Cross-interaction hierarchical attention networks for urban anomaly prediction. In *IJCAI*.

[Li et al. 2018] Li, Y.; Yu, R.; Shahabi, C.; and Liu, Y. 2018. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *ICLR*.

[Liang et al. 2018] Liang, Y.; Ke, S.; Zhang, J.; Yi, X.; and Zheng, Y. 2018. Geoman: Multi-level attention networks for geo-sensory time series prediction. In *IJCAI*, 3428–3434.

[Liang et al. 2019] Liang, Y.; Ouyang, K.; Jing, L.; Ruan, S.; Liu, Y.; Zhang, J.; et al. 2019. Urbanfm: Inferring fine-grained urban flows. In *KDD*, 3132–3142. ACM.

[Liu et al. 2016] Liu, Q.; Wu, S.; Wang, L.; et al. 2016. Predicting the next location: A recurrent model with spatial and temporal contexts. In *AAAI*, 194–200.

[Pan et al. 2019] Pan, Z.; Liang, Y.; Wang, W.; et al. 2019. Urban traffic prediction from spatio-temporal data using deep meta learning. In *KDD*. ACM.

[Pan, Demiryurek, and others 2012] Pan, B.; Demiryurek, U.; et al. 2012. Utilizing real-world transportation data for accurate traffic prediction. In *ICDM*, 595–604. IEEE.

[Shen et al. 2018] Shen, B.; Liang, X.; Ouyang, Y.; Liu, M.; Zheng, W.; and Carley, K. M. 2018. Stepdeep: a novel spatial-temporal mobility event prediction framework based on deep neural network. In *KDD*, 724–733.

[Song et al. 2020] 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.

[Srinivasan, Chan, and Balaji 2009] Srinivasan, D.; Chan, C. W.; and Balaji, P. 2009. Computational intelligence-based congestion prediction for a dynamic urban street network. *Neurocomputing* 72(10-12):2710–2716.

[Wang and Li 2017] Wang, H., and Li, Z. 2017. Region representation learning via mobility flow. In *CIKM*, 237–246.

[Wang et al. 2020] Wang, X.; Ma, Y.; Wang, Y.; Jin, W.; Wang, X.; et al. 2020. Traffic flow prediction via spatial temporal graph neural network. In *WWW*, 1082–1092.

[Wei et al. 2018] Wei, H.; Zheng, G.; Yao, H.; and Li, Z. 2018. Intellilight: A reinforcement learning approach for intelligent traffic light control. In *KDD*, 2496–2505.

[Wu et al. 2018] Wu, X.; Shi, B.; Dong, Y.; Huang, C.; Faust, L.; and Chawla, N. V. 2018. Restful: Resolution-aware forecasting of behavioral time series data. In *CIKM*, 1073–1082.

[Yao et al. 2018] Yao, H.; Wu, F.; Ke, J.; Tang, X.; Jia, Y.; Lu, S.; et al. 2018. Deep multi-view spatial-temporal network for taxi demand prediction. In *AAAI*, 2588–2595.

[Yao et al. 2019] 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*.

[Yu et al. 2017a] Yu, R.; Li, Y.; Shahabi, C.; et al. 2017a. Deep learning: A generic approach for extreme condition traffic forecasting. In *SDM*, 777–785. SIAM.

[Yu et al. 2017b] Yu, R.; Zheng, S.; Anandkumar, A.; and Yue, Y. 2017b. Long-term forecasting using tensor-train rnns. *Arxiv*.

[Yu, Yin, and Zhu 2018] Yu, B.; Yin, H.; and Zhu, Z. 2018. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *IJCAI*.

[Zhang et al. 2016] Zhang, J.; Zheng, Y.; Qi, D.; Li, R.; and Yi, X. 2016. Dnn-based prediction model for spatio-temporal data. In *SIGSPATIAL*, 1–4.

[Zhang et al. 2020] Zhang, X.; Huang, C.; Xu, Y.; et al. 2020. Spatial-temporal convolutional graph attention networks for citywide traffic flow forecasting. In *CIKM*, 1853–1862.

[Zhang, Zheng, and Qi 2017] Zhang, J.; Zheng, Y.; and Qi, D. 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *AAAI*.

[Zhao et al. 2017] Zhao, L.; Sun, Q.; Ye, J.; Chen, F.; Lu, C.-T.; and Ramakrishnan, N. 2017. Feature constrained multi-task learning models for spatiotemporal event forecasting. *TKDE* 29(5):1059–1072.

[Zheng et al. 2020] Zheng, C.; Fan, X.; Wang, C.; and Qi, J. 2020. Gman: A graph multi-attention network for traffic prediction. In *AAAI*, 1234–1241.