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]

[1]South China University of Technology, China
[2]JD Finance America Corporation, USA
[3]Communication and Computer Network Laboratory of Guangdong, China
[4]Peng Cheng Laboratory, China
[5]JD Intelligent Cities Research, China, [6]JD Intelligent Cities Business Unit, 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. First, only the neighboring spatial correlations among adjacent regions are considered in most existing methods, and the global inter-region dependency is ignored. Additionally, these methods fail to encode the complex traffic transition regularities exhibited with time-dependent and multi-resolution in nature. To tackle these challenges, we develop a new traffic prediction framework–**S**patial-**T**emporal **G**raph **D**iffusion **N**etwork (ST-GDN). 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 several real-life traffic datasets demonstrate that ST-GDN outperforms different types of state-of-the-art baselines. Source codes of implementations are available at *https://github.com/jill001/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).

With 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. 2017). 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 the temporal dimension (Yao et al. 2018). Although both spatial and temporal correlations have been considered in existing methods, several key challenges have not been well addressed.

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

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 global relations across regions,

which leads to suboptimal predictions results.

To tackle the above challenges, we propose a new predictive framework **S**patial-**T**emporal **G**raph **D**iffusion **N**etwork (ST-GDN), for region-specific traffic flow. In ST-GDN, we develop a multi-scale self-attention network to investigate multi-grained temporal dynamics across various time resolutions, in order to encode temporal hierarchy of traffic transitional regularities. To promote the collaboration of different granularity-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 of incorporating spatial semantics from local-level spatially adjacent relations to global-level traffic pattern representations across the city in a joint way.

We highlight the key contributions of this work as below:

- 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 granularity-aware latent representations, with the cooperation of the designed multi-scale self-attention network and temporal hierarchy aggregation layer.

- ST-GDN preserves both local and global region-wise dependencies, via a hierarchically structured graph neural architecture which is consisted of a graph attention network and convolution-based graph diffusion mechanism.

- Our extensive experiments on three 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 in the traffic prediction process.

Problem Definition

In this section, we begin with some key definitions and preliminary terms which are relevant to the solution.

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 future traffic volume of regions.

Methodology

This section presents the our ST-GDN with the descriptions of different components (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:

$$\mathbf{Q} = \mathbf{E}^p \cdot \mathbf{W}_Q, \mathbf{K} = \mathbf{E}^p \cdot \mathbf{W}_K, \mathbf{V} = \mathbf{E}^p \cdot \mathbf{W}_V. \quad (1)$$

We further perform the operation of $\mathbf{Y}^p = \sigma(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V}$, 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. Here, $\mathbf{W}_Q$, $\mathbf{W}_K$, $\mathbf{W}_V$ are projection matrices.

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-based 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.

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

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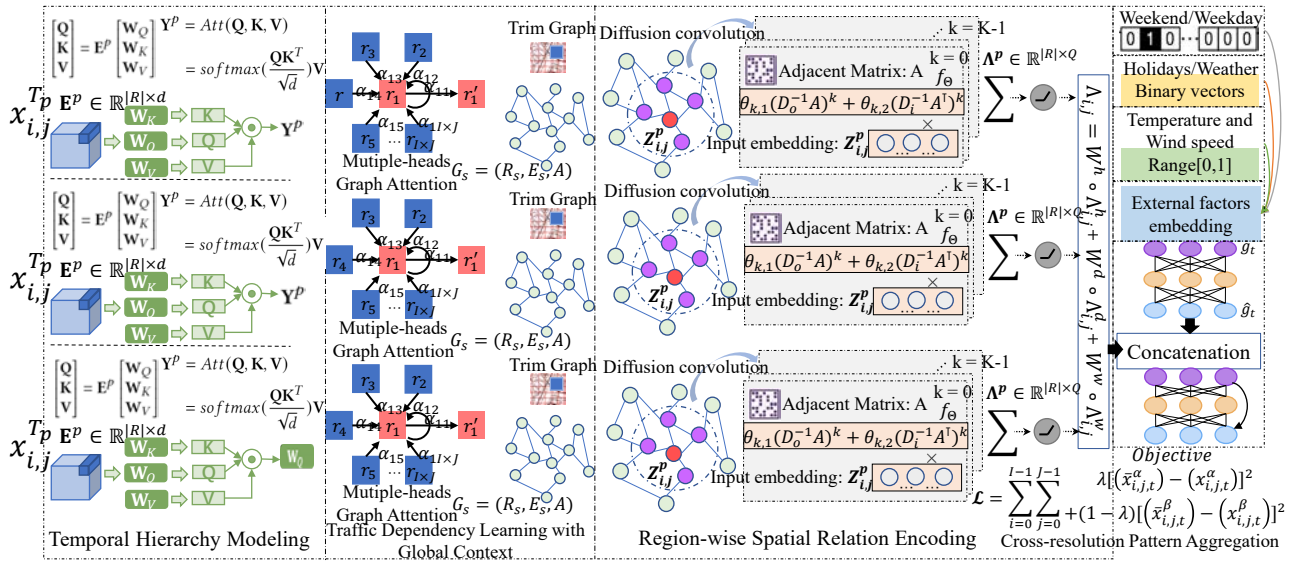
Figure 1: The framework of our developed spatial-temporal graph diffusion networks.

$\omega^h_{(i,j);(i',j')}$ 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 α 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:

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

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:

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

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, motivated by (Li et al. 2018), 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 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 nodes in G_s to generate new feature representations:

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

where $\theta_{k,1}$, $\theta_{k,2}\in\mathbb{R}^{K\times2}$. $D_o^{-1}\mathbf{A}$ (in-degree) and $D_i^{-1}\mathbf{A}^\mathsf{T}$ (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 $\Theta\in\mathbb{R}^{Q\times d\times K\times2}$, in which the Q-dimensional output $\boldsymbol{\Lambda}^p\in\mathbb{R}^{|R|\times Q}$ of diffusion convolutional layer is given:

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

where $q\in\{1,...,Q\}$. The obtained region representation $\boldsymbol{\Lambda}^p_{i,j}$ 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 $\boldsymbol{\Lambda}^p_{i,j}$ 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 ($\boldsymbol{\Lambda}^{p_h}$), daily ($\boldsymbol{\Lambda}^{p_d}$) and weekly ($\boldsymbol{\Lambda}^{p_w}$) as: $\boldsymbol{\Lambda}_{i,j} = \mathbf{W}^h\circ\boldsymbol{\Lambda}^h_{i,j} + \mathbf{W}^d\circ\boldsymbol{\Lambda}^d_{i,j} + \mathbf{W}^w\circ\boldsymbol{\Lambda}^w_{i,j}$. 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 $\boldsymbol{\Lambda}_{i,j}$ which preserves multi-grained temporal hierarchy of traffic regularities.

Traffic Prediction Phase

In urban sensing, 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 external factors. In particular, we consider several types of external factors: Weather conditions, Temperature/°C, Wind speed/mph. We map 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 of each region.

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:

$$\mathcal{L} = \sum_{i=0}^{I-1} \sum_{j=0}^{J-1} \lambda[(\bar{x}_{i,j,t}^{\alpha}) - (x_{i,j,t}^{\alpha})]^2 \qquad (7)$$
$$+ (1-\lambda)[(\bar{x}_{i,j,t}^{\beta}) - (x_{i,j,t}^{\beta})]^2$$

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

Model Complexity Analysis. In this part, 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 attentional graph 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 component takes $O(K \times |E_s|)$ 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. Experiments are performed on three 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 32×32 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 10×20 grid map. The traffic data sample period is also half an hour.
NYC-Bike (Zhang, Zheng, and Qi 2017). It includes the trajectories of the bike system from New York with a 16×8 grid map. Traffic volume is estimated on an hourly basis.

Evaluation Protocols. We leverage two representative metrics for evaluation: *Root Mean Squared Error (RMSE)* and *Mean Absolute Percentage Error (MAPE)*.

Methods for Comparison. In the comparison, we consider the following baselines with various model structures.

- **ARIMA** (Pan, Demiryurek et al. 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.

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

- **RNN** (Liu et al. 2016): it leverages the recurrent neural networks for capturing both the spatial and temporal effects for making sequential data prediction.

- **LSTM** (Yu et al. 2017): it jointly models the normal and abnormal traffic variations based on stacked long short-term memory networks.

- **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.

- **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.

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

- **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.

- **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.

Dataset	BJ-Taxi	NYC-Taxi	NYC-Bike	External Factors	Beijing	New York City
Data type	Taxi GPS	Taxi GPS	Bike Rental	Weather	sunny, rainy *et al.*	sunny, rainy *et al.*
Time interval	30 minutes	30 minutes	one hour	Temperature/°C	$[-24.6, 41.0]$	$[-10.3, 31.40]$
Gird map size	32×32	10×20	16×8	Wind speed/mph	$[0, 48.60]$	$[0, 63.75]$
# of records	34,000+	22,000,000+	6,800+	holidays	weekends, holidays	weekends, holidays

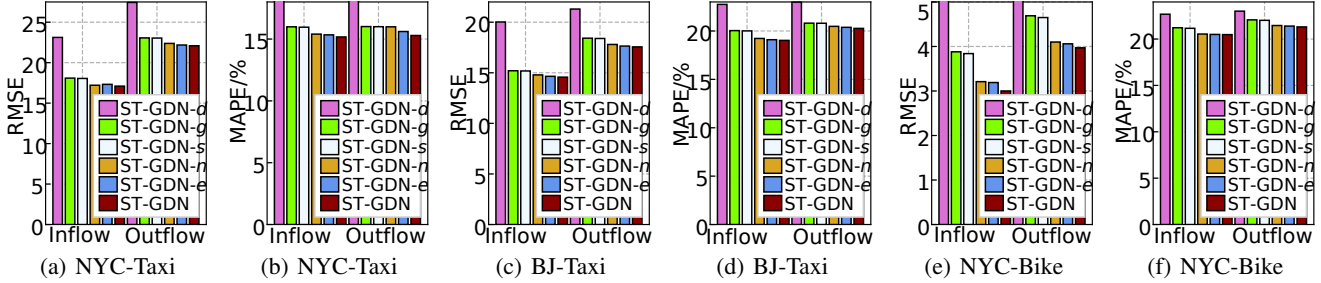Table 1: Statistical information of experimented datasets.

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

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 comparison results of all methods are presented in Table 2. 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 2.

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.
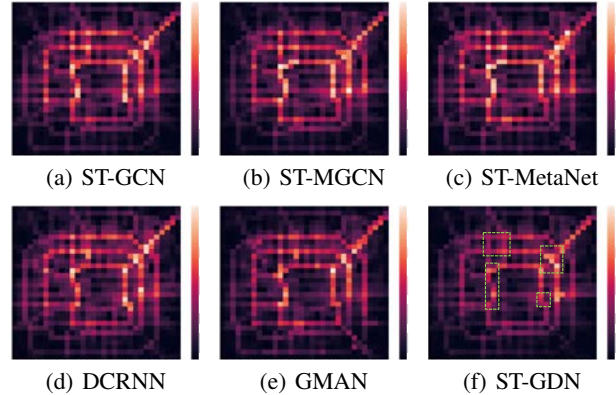

Figure 3: Visualization for Traffic Prediction Errors.

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.

- *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

Datasets	BJ-Taxi				NYC-Bike				NYC-Taxi			
Metrics	RMSE		MAPE (%)		RMSE		MAPE (%)		RMSE		MAPE (%)	
Methods	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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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**

Table 2: Performance comparison of all methods on three datasets in terms of *RMSE* and *MAPE*.

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.

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.

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 efficiency (measured by running time) of our *ST-GDN*. Table 3 presents the computational cost of training (with 300 epochs) and inference phase for *ST-GDN* and five best performed baselines on three 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.
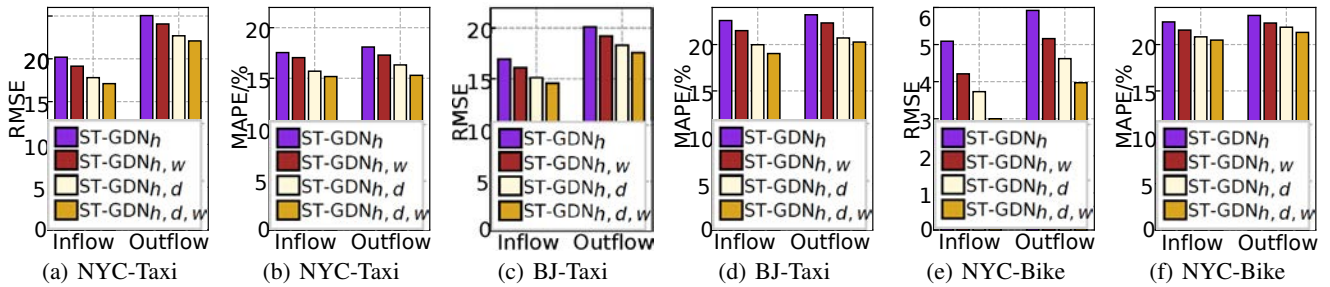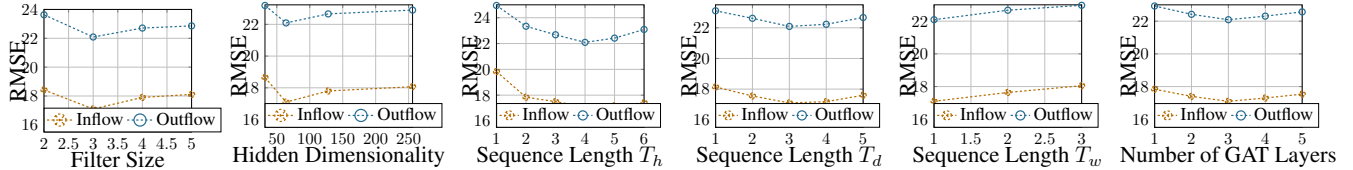
Figure 4: Multi-resolution temporal effect studies.

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

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

Table 3: Model Efficiency Study.

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. 2017; 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 based on convolution neural network. In this work, ST-GDN incorporates the global context enhanced region-wise explicit relevance into a graph diffusion paradigm to capture comprehensive high-order region dependencies in a joint learning manner.

Graph-based Spatial-Temporal Prediction. It is worth mentioning that several recent efforts have investigated Graph Neural Networks (GNNs) for spatial-temporal data forecasting (Guo et al. 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 cross-region inter-correlations, 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 many baselines over several datasets consistently. Our future work lies in the deployment of our developed prototype in a cloud-based working system for real-time traffic prediction.

Acknowledgments

References

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

Deng, D.; Shahabi, C.; Demiryurek, U.; Zhu, L.; Yu, R.; and Liu, Y. 2016. Latent space model for road networks to predict time-varying traffic. In *KDD*, 1525–1534.

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, 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, X.; Li, Y.; Wang, L.; Zhang, L.; Yang, Q.; Ye, J.; and Liu, Y. 2019. Spatiotemporal Multi-Graph Convolution Network for Ride-hailing Demand Forecasting. In *AAAI*.

Guo, S.; Lin, Y.; Feng, N.; Song, C.; and Wan, H. 2019. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *AAAI*, volume 33, 922–929.

Huang, C.; Zhang, C.; Dai, P.; and Bo, L. 2020. Cross-Interaction Hierarchical Attention Networks for Urban Anomaly Prediction. In *IJCAI*.

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, C.; Zhang, J.; Zheng, Y.; and Chawla, N. V. 2018. DeepCrime: attentive hierarchical recurrent networks for crime prediction. In *CIKM*, 1423–1432.

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

Liang, Y.; Ouyang, K.; Jing, L.; Ruan, S.; Liu, Y.; Zhang, J.; Rosenblum, D. S.; and Zheng, Y. 2019. UrbanFM: Inferring Fine-Grained Urban Flows. In *KDD*, 3132–3142. ACM.

Liu, Q.; Wu, S.; Wang, L.; and Tan, T. 2016. Predicting the Next Location: A Recurrent Model with Spatial and Temporal Contexts. In *AAAI*, 194–200.

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

Pan, Z.; Liang, Y.; Wang, W.; Yu, Y.; Zheng, Y.; and Zhang, J. 2019. Urban Traffic Prediction from Spatio-Temporal Data Using Deep Meta Learning. In *KDD*. ACM.

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

Wang, X.; Ma, Y.; Wang, Y.; Jin, W.; Wang, X.; Tang, J.; Jia, C.; and Yu, J. 2020. Traffic Flow Prediction via Spatial Temporal Graph Neural Network. In *WWW*, 1082–1092.

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, 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, H.; Tang, X.; Wei, H.; Zheng, G.; and Li, Z. 2019. Revisiting Spatial-Temporal Similarity: A Deep Learning Framework for Traffic Prediction. In *AAAI*.

Yao, H.; Wu, F.; Ke, J.; Tang, X.; Jia, Y.; Lu, S.; Gong, P.; Ye, J.; and Li, Z. 2018. Deep multi-view spatial-temporal network for taxi demand prediction. In *AAAI*, 2588–2595.

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

Yu, R.; Li, Y.; Shahabi, C.; Demiryurek, U.; and Liu, Y. 2017. Deep learning: A generic approach for extreme condition traffic forecasting. In *SDM*, 777–785. SIAM.

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

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, X.; Huang, C.; Xu, Y.; et al. 2020. Spatial-Temporal Convolutional Graph Attention Networks for Citywide Traffic Flow Forecasting. In *CIKM*, 1853–1862.

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