Spatial-Temporal Convolutional Graph Attention Networks for Citywide Traffic Flow Forecasting

Xiyue Zhang
South China University of Technology
zhang.xiyue@mail.scut.edu.cn

Chao Huang
JD Finance America Corporation
chaohuang75@gmail.com

Yong Xu*
South China University of Technology
Peng Cheng Laboratory
yxu@scut.edu.cn

Lianghao Xia
South China University of Technology
cslianghao.xia@mail.scut.edu.cn

ABSTRACT

Traffic flow prediction plays an important role in many spatial-temporal data applications, *e.g.*, traffic management and urban planning. Various deep learning techniques are developed to model the traffic dynamic patterns with different neural network architectures, such as attention mechanism, recurrent neural network. However, two important challenges have yet to be well addressed: (i) Most of these methods solely focus on local spatial dependencies and ignore the global inter-region dependencies in terms of traffic distributions; (ii) It is important to capture channel-aware semantics when performing spatial-temporal information aggregation. To address these challenges, we propose a new traffic prediction framework–**S**patial-**T**emporal **C**onvolutional **G**raph **A**ttention Network (ST-CGA), to enable the traffic prediction with the modeling of region dependencies, from locally to globally in a comprehensive manner. In our ST-CGA framework, we first develop a hierarchical attention networks with a graph-based neural architecture, to capture both the multi-level temporal relations and cross-region traffic dependencies. Furthermore, a region-wise spatial relation encoder is proposed to supercharge ST-CGA mapping spatial and temporal signals into different representation subspaces, with channel-aware recalibration residual network. Extensive experiments on four real-world datasets demonstrate that ST-CGA achieve substantial gains over many state-of-the-art baselines. Source codes are available at: https://github.com/shurexiyue/ST-CGA.

CCS CONCEPTS

• **Information systems → Spatial-temporal systems**; **Data mining**; • **Computing methodologies → Neural networks**;

KEYWORDS

Spatial-temporal Data Mining; Graph Attention Networks; Traffic flow Forecasting, Urban Computing

ACM Reference Format:
Xiyue Zhang, Chao Huang, Yong Xu, and Lianghao Xia. 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 (CIKM '20), October 19–23, 2020, Virtual Event, Ireland.* ACM, Taipei, 10 pages. https://doi.org/10.1145/3340531.3411941

1 INTRODUCTION

Spatial-temporal forecasting has become a key component in mining geographical data across time and space dimensions, such as ride-hailing demand management [3], geo-anomaly prediction [9] and user location prediction [16]. Among different spatial-temporal mining applications, forecasting traffic flow (*e.g.*, input and output traffic volume of each geographical region in a city) is very important for the information infrastructure systems of smart city applications [5]. For example, accurate predictions of citywide traffic flow information for the next time slot could help the government for making decisions about transportation management (*e.g.*, enhance the running efficiency of transportation hub), and improve the public safety risk assessment (*e.g.*, mitigate the tragedies caused by the crowd flow) [18].

As the related work stated, the temporal traffic variation patterns and the underlying spatial dependencies across regions have high impacts on the future traffic information. Hence, many neural network models have been proposed to learn the spatial and temporal correlations based on various deep learning frameworks, such as the convolutional recurrent network [34], attentive recurrent model [28] and graph neural network [30]. Despite the effectiveness of existing traffic flow forecasting methods, we identify three significant challenges that have not been addressed well.

First, while spatial correlations between regions are considered in existing research literature, most of them only model the local dependencies with the consideration of nearby geographical relationships [29]. However, two regions can also be correlated with each other in terms of traffic distribution, even though they are not spatially adjacent in the urban area [3]. In such cases, when modeling the traffic evolving patterns, we need to consider the region-wise dependencies in a global space. Otherwise, it is likely

*Corresponding author

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3411941

that the learned relations among geographical regions are inaccurate, and thus the forecasting performance of future citywide traffic information is suboptimal. *Second*, existing works have modeled citywide traffic as a heatmap image and use convolutional neural network to model the non-linear spatial correlations [34]. While the position-based dependencies between regions can be encoded by the convolution-based kernels with latent channels, different channel-specific representation subspaces are equally treated. During the spatial pattern integration, as the importance of different channel views can be quite different corresponding to different spatial-temporal semantics, their weights in assisting future traffic prediction need to be carefully decided.

Motivated by the aforementioned challenges, we propose a new traffic flow forecasting framework–**S**patial-**T**emporal **C**onvolutional **G**raph **A**ttention Network (ST-CGA). Specifically, this work first proposes a hierarchically structured attention network, to jointly learn granularity-specific embeddings by preserving multi-level temporal patterns and global cross-region traffic dependencies, based on an integrative framework of the self-attention layer and graph attention network. By stacking multiple granularity-aware hierarchical encoders, we can inject multi-granularity intra-region temporal signals into the inter-region relation learning. In addition, to fully capture spatial semantics across latent channel dimensions, we propose the channel-aware recalibration network to learn the weights of different channels during spatial-temporal pattern aggregation. Finally, a cross-granularity gating mechanism is introduced to enable the fusion of different temporal representations of individual time granularity. We also develop an external factor fusion module to capture the influences of external data sources (*e.g.*, meteorological data) in the prediction phase.

The main contributions of this work are highlighted as follows:

- This work explicitly explores the multi-granularity temporal dynamics and global region-wise dependencies, under channel-aware representation learning spaces, in studying the traffic prediction problem.
- We develop a hierarchically structured attention networks with the integrative framework of self-attention module and graph attention subnet, to jointly model the multi-granularity temporal dynamics and global inter-region traffic dependencies.
- We propose a new multi-view collaboration framework which explicitly embeds multi-level temporal signals into the channel-aware spatial relation encoder, with the cooperation of the designed convolution-based recalibrated residual network and cross-granularity gating mechanism.
- We validate our framework over four real-world traffic datasets to demonstrate that ST-CGA outperforms different types of baselines in yielding better prediction performance. We further perform case study with qualitative examples to better understand the interpretation of ST-CGA.

The remaining part of this paper is organized as follows. We introduce key definitions and problem formulation in Section 2. Section 3 describes the technical details of our ST-CGA framework. We report the evaluation results in Section 4 to show the effectiveness of our ST-CGA method. Section 5 discusses the related work. Finally, this work is concluded in Section 6.

2 PRELIMINARY

We first introduce some preliminary terms and then formally present the problem statement. For the notations, we use the lowercase bold for vectors, and capital bold for matrices and tensors.

Definition 1. ***Geographical Region***. *We divide a city into $M \times N$ disjoint regions based on geographical coordinate information (latitude and longitude). We define each partitioned grid as a geographical region $r_{m,n}$, where m and n is the index of M and N, respectively. Each region $r_{m,n} \in R$ is our target spatial unit for traffic flow prediction, where R is defined as the set of all partitioned regions.*

Based on the partitioned geographical regions, we further define region's traffic flow information over time as follows:

Definition 2. ***Traffic Flow Tensor***. *We define two three-way tensors $\mathcal{X}^i \in \mathbb{R}^{M \times N \times T}$ and $\mathcal{X}^o \in \mathbb{R}^{M \times N \times T}$ to represent the inflow and outflow of each region. In particular, the element $x^i_{m,n,t}$ in $\mathcal{X}^i$ represents the incoming traffic volume (e.g., the number of vehicle arrivals) of region $r_{m,n}$ at the t-th time interval (e.g., hour or day). Each entry $x^o_{m,n,t}$ of $\mathcal{X}^o$ denotes the outgoing traffic volume (e.g., the number of departed vehicles) of region $r_{m,n}$ at the t-th time interval.*

Problem Statement. The objective of citywide traffic flow prediction is that: given observations of historical traffic flow information (*i.e.*, $\mathcal{X}^i_T \in \mathbb{R}^{M \times N \times T}$ and $\mathcal{X}^o_T \in \mathbb{R}^{M \times N \times T}$) from previous T time intervals across all regions ($r_{m,n}$, $(m,n) \in [(1,1), ..., (M,N)]$) in a city, we aim to learn a predictive function which infers the unknown traffic flow volume in future time intervals.

3 METHODOLOGY

In this section, we elaborate our ST-CGA model with the technical details. The model architecture is shown in Figure 1.

3.1 Inter-Region Traffic Dependency Encoding

Due to the multi-periodic patterns of traffic flow variation and complex inter-region traffic correlations, we propose to capture the dependencies across all regions in terms of their traffic flow distributions with the consideration of different time granularities. Given the defined traffic flow tensor $\mathcal{X}^i \in \mathbb{R}^{M \times N \times T}$ and $\mathcal{X}^o \in \mathbb{R}^{M \times N \times T}$, we first respectively generate the corresponding multiple granularity-aware traffic follow tensor for the inflow and outflow of regions. Without loss of generality and for simplifying notations, we do not differentiate the traffic inflow and outflow by unity their individual data point as $x_{m,n,t}$, where (m,n) and t are region and time identifiers, respectively. We define the following terms to be used in our predictive method.

In our ST-CGA framework, motivated by the work [23, 35], we propose to consider multiple time granularities (*e.g.*, hour, day and week) in model the temporal patterns of traffic volume across time dimension. In particular, we regard each time granularity as the period to sample the traffic data point and generate the corresponding granularity-specific traffic data series. For example, given the target granularity is hour, the time difference between the sampled two traffic volume measurements $x_{m,n,t}$ and $x_{m,n,t'}$ is a hour.

In this component, we generate different granularity-aware traffic flow tensor based on the original input tensor $\mathcal{X}$ corresponding to different settings of time granularities.

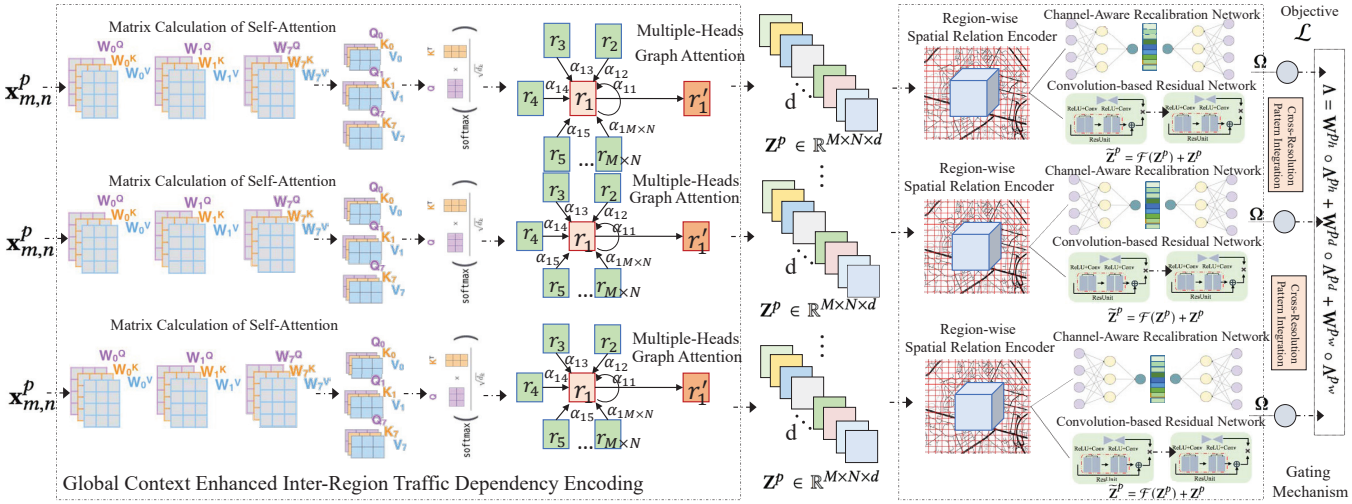


Figure 1: The framework of our developed spatial-temporal convolutional graph attention networks.

Definition 3. ***Granularity-aware Traffic Flow Tensor*** $\mathcal{X}^p$. *With the different period granularity settings, we define tensor* $\mathcal{X}^p \in \mathbb{R}^{M\times N\times T_p}$ *to represent granularity-specific traffic flow data of region* $r_{m,n}$ *with the time granularity of p (e.g., hour, day), where* T_p *denotes the length of input traffic series* $X^p_{m,n}$ *and* T_p *can vary by different settings of p. Given the most fine-grained sample time granularity is hour, we take the* $p \in \{hour, day\}$ *as concrete examples and generate the corresponding region-specific traffic volume vector as follows:*

$$\mathbf{x}^p_{m,n} = (x^{p,1}_{m,n}, ...x^{p,(t-1)}_{m,n}, x^{p,t}_{m,n}, x^{p,(t+1)}_{m,n}, ..., x^{p,T_p}_{m,n})(p = hour)$$
$$\mathbf{x}^p_{m,n} = (x^{p,1}_{m,n}, ...x^{p,(t-24)}_{m,n}, x^{p,t}_{m,n}, x^{p,(t+24)}_{m,n}, ..., x^{p,T_p}_{m,n})(p = day) \quad (1)$$

3.1.1 **Granularity-Aware Self-Attention Encoder.** Based on the constructed granularity-specific traffic volume vector $\mathbf{x}^p_{m,n}$ of region $r_{m,n}$, we develop a self-attention layer to encode the traffic evolving patterns, by automatically performing the temporal aggregation over vector $\mathbf{x}^p_{m,n}$. Self-attention has been demonstrated the effectiveness in various sequence learning applications, such as nature language processing [25, 26] and user behavior modeling [12, 24]. With the format of matrix calculation in self-attention layer, we first define the query, key and value matrices ($\mathbf{Q} \in \mathbb{R}^{T_p\times d}$, $\mathbf{K} \in \mathbb{R}^{T_p\times d}$ and $\mathbf{V} \in \mathbb{R}^{T_p\times d}$), to project the input traffic information $\mathcal{X}^p \in \mathbb{R}^{T_p\times d}$ across all regions into three dimensions of latent representations which can be used for self-attention weights calculations. To generate these three vectors for each region-wise traffic flow, we perform linear transformation on the input embedding matrix $\mathbf{E}^p$ with the period granularity of p across all regions (each row corresponds to an individual region $r_{m,n}$) with the learnable weight matrices $\mathbf{W}_Q$, $\mathbf{W}_K$ and $\mathbf{W}_V$, corresponding to the construction of the query, key and value matrices, respectively. We further formally present the self-attention layer with the dot-product attention operation as:

$$\begin{bmatrix}\mathbf{Q}\\ \mathbf{K}\\ \mathbf{V}\end{bmatrix} = \mathbf{E}^p \begin{bmatrix}\mathbf{W}_Q\\ \mathbf{W}_K\\ \mathbf{W}_V\end{bmatrix};\ \mathbf{Y}^p = Att(\mathbf{Q},\mathbf{K},\mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V} \quad (2)$$

where $\mathbf{Y}^p \in \mathbb{R}^{|R|\times d}$ represents the encoded embeddings from the temporally-ordered traffic information with the period granularity of p. Each row $\mathbf{y}^p_{m,n}$ in $\mathbf{Y}^p$ corresponds to individual region $r_{m,n}$.

3.1.2 **Region-wise Graph Attention Module.** After obtaining the learned granularity-specific representation $\mathbf{y}^p_{m,n}$ of each region $r_{m,n}$, we propose a graph attention network to capture the complex inter-dependencies across different geographical regions with respect to their traffic volume distributions. To be specific, we first define a region graph $G = (V, E)$, where V and E represents the set of vertices and edges, respectively. In graph G, each vertex indicates an individual region $r_{m,n}$ and each edge represents the pairwise relationship between two geographical regions. Based on the initialized embeddings $\mathbf{Y}^p$ (with the period granularity of p, such as hour or day) encoded from the granularity-aware self-attention layer, we perform the information aggregation with the graph-based attention encoder to preserve the high-order region-wise traffic relations from a global perspective. The general idea of graph-structured attentive relation encoding is to learn which regions are able to attend in terms of their traffic patterns. It addresses the limitation of graph convolutions in differentiating dependencies across regions [20], *i.e.*, graph convolutional network is a structure-dependent message aggregation architecture and can hardly explicitly capture the different pairwise relevance between instance nodes in the graph.

The granularity-aware graph attention module of our ST-CGA architecture is a four-phase framework. To enhance the expressive power of feature representation during the graph-based aggregation process, we first perform linear transformation on the input feature vector of each region $\mathbf{y}^p_{m,n}$ in $\mathbf{Y}^p$, with a shared parametrized weight matrix $\mathbf{W}_p$ based on the following formal operation:

$$\widetilde{\mathbf{Y}}^p = \mathbf{Y}^p \cdot \mathbf{W}_p,\quad \mathbf{W}_p \in \mathbb{R}^{d\times d'} \quad (3)$$

where d' is the embedding dimension of high-level feature representations after the projection operation. Then, we aim to compute a pairwise attention coefficient $\epsilon_{(m,n),(m',n')}$ between the neighboring region $r_{m,n}$ and $r_{m',n'}$. We concatenate the projected embedding $\widetilde{\mathbf{y}}^p_{m,n}$ and $\widetilde{\mathbf{y}}^p_{m',n'}$ and take a dot product of them with a trainable weight vector $\boldsymbol{\alpha}$. The activation function of LeakyReLU and the

softmax are further applied to generate the attention coefficient $\boldsymbol{\alpha}$, which is formally represented as follows:

$$\epsilon_{(m,n),(m',n')} = \frac{exp(LR(\boldsymbol{\alpha}^T[\widetilde{\mathbf{y}}^p_{m,n}||\widetilde{\mathbf{y}}^p_{m',n'}]))}{\sum_{(m',n')\in\mathcal{N}(m,n)} exp(LR(\boldsymbol{\alpha}^T[\widetilde{\mathbf{y}}^p_{m,n}||\widetilde{\mathbf{y}}^p_{m',n'}]))} \quad (4)$$

where $||$ denotes the concatenation operation and $LR(\cdot)$ represents the *LeakyReLU* function. Analogous to multiple representation channels in the convolutional neural architecture, we stabilize the learning process of our graph attention module by employing the multi-head attention mechanism with the following operation, to update the feature embedding $\widetilde{\mathbf{y}}^p_{m,n}$ of each region $r_{m,n}$:

$$\mathbf{z}^p_{m,n} = ||_{h=1}^{H} LeakyReLU\Big(\sum_{(m',n')\in\mathcal{N}(m,n)} \epsilon^h_{(m,n),(m',n')}\widetilde{\mathbf{y}}^p_{m,n}\Big) \quad (5)$$

where $\mathbf{z}^p_{m,n} \in \mathbf{Z}^p$ is the refined feature representation of region $r_{m,n}$ with the preservation of cross-region dependencies in terms of their traffic distributions (with the period granularity of p). We define the updated feature representations as $\mathbf{Z}^p \in \mathbb{R}^{M\times N\times d}$ of all regions ($r_{m,n} \in R$). $\epsilon^h_{(m,n),(m',n')}$ to denote the attentive coefficient of the h-th head learning space. Here, H denotes the total number of heads. With the design of our dual-stage attention architecture in ST-CGA, we can capture the rich semantics behind region's traffic dynamic correlations across the entire urban space.

3.2 Spatial Relation Modeling between Regions

In this subsection, we propose to encode the spatial relationships between regions across different latent channel dimensions, with the designed convolution-based recalibrated residual network.

3.2.1 **Convolution-based Residual Unit.** Based on the encoded latent representations $\mathbf{Z} \in \mathbb{R}^{M\times N\times d\times|P|}$ of region traffic transitional regularities, we further propose a convolution-based residual module to model the spatial relation structures between different geographical regions. Specifically, we feed $\mathbf{Z}$ into an integrative architecture which is composed of convolution operation and residual unit. In our spatial relation encoder, we employ the ResNet for each individual granularity p with the residual mappings, to alleviate the vanishing-gradient issue and strengthen feature propagation [6]. Formally, the residual unit is defined as below:

$$\widetilde{\mathbf{Z}}^p = \mathcal{F}(\mathbf{Z}^p) + \mathbf{Z}^p \quad (6)$$

where $\mathcal{F}(\cdot)$ is a residual operator and $\widetilde{\mathbf{Z}}^p$ denoted the encoded high-level feature representation with the period granularity of p. The residual operator $\mathcal{F}(\cdot)$ is composed of two stacked convolutional layers, which is formally defined as follows:

$$\mathbf{Z}^{p(l+1)} = ReLU(W^{(l)} * \mathbf{Z}^{p(l)} + b^{(l)}) \quad (7)$$

$W^{(l)}$ and $b^{(l)}$ are learnable parameters. $\mathbf{Z}^{p(l+1)} \in \mathbb{R}^{M\times N\times C}$ is the encoded representation vector after the Conv+ReLU operation, where C is the number of filters. $*$ indicates the convolutional operation. The *ReLU* function is applied to the convolution layer. In our convolutional network, the kernel size is set as 3×3 with the stride of 1. The final obtained $\widetilde{\mathbf{Z}}^p \in\in \mathbb{R}^{M\times N\times C}$ further injects the geographical contextual signals after the traffic dependency modeling across regions.

3.2.2 **Channel-Aware Recalibration Network.** Motivated by the feature learning paradigm with hierarchical network structure [15], we further develop a channel-aware recalibration network with a top-down and bottom-up architecture, to augment our spatial relation encoder by capturing the relationships across different representation channels. To achieve this goal, we propose to learn a mask tensor $\Omega \in \mathbb{R}^{M\times N\times C}$ to represent the different importance scores along with channel dimensions. We examine two candidate encoder functions in our recalibration network, and their effects are investigated in our experiments of Section 4.6.

Fully Convolutional Networks (FConv). Our first candidate encoder is to perform convolution and max pooling operations alternately, to increase the receptive field and obtain the intermediate hidden representation with a top-down architecture. Then, the global feature interaction signals across all regions and channels, is then expanded by a symmetrical top-down architecture to generate the weights of the input for each position in $\mathbf{Z}$. Linear interpolation up samples the output after convolutions. The number of bi-linear interpolation is the same as that of max pooling to be consistent with the input embedding dimensionality. The final output is constrained to $[0, 1]$ with the normalization to get the final mask Ω.

Fully Connected Layers (FCL). Another encoder function is to stretch the input representation tensor $\mathbf{Z}^p$ into a one-dimensional vector, and feed it into a stacked feed-forward neural networks to generate an intermediate latent representation. A symmetric structure of bottom-up network with a reverse order of stacked feed-forward networks, is utilized for learning the mask tensor Ω. Similarly, a sigmoid function is applied to map the output into the range of $[0, 1]$ to generate the final mask Ω.

With the joint consideration of region spatial relations across $\mathbb{R}^{M\times N}$ and channels $\mathbb{R}^C$, we could generate the mask tensor Ω– corresponding to each element position in $\widetilde{\mathbf{Z}}^p$. We further apply the mask tensor on granularity-specific representation $\widetilde{\mathbf{Z}}^p$, to obtain Λ^p with the following recalibration operation:

$$\boldsymbol{\Lambda}^p = \Omega \circ \widetilde{\mathbf{Z}}^p = \Omega \circ \mathcal{F}(\mathbf{Z}^p) + \mathbf{Z}^p \quad (8)$$

where $\circ$ is the element-wise multiplication. By integrating the channel-aware recalibration network into ST-CGA with the learned relevance scores into the traffic pattern representation process, the generated mask tensor Ω i) enhances the representation capability by explicitly differentiating dimension units (*i.e.*, regions and latent channels); ii) serves as a gradient update filter during the back propagation, *i.e.*, enhancing the robustness of *ST-CGA* in learning gradients for parameter inference in convolutional residual unit [21].

3.3 Cross-granularity Pattern Integration

To aggregate the complex spatial and temporal patterns (encoded from the architecture of channel-aware convolutional graph attention network), we develop a gating mechanism to enable the fusion of different granularity-specific representations $\boldsymbol{\Lambda}^p$ ($p \in P = \{hour, day, week\}$). Each one corresponds to the hourly ($\boldsymbol{\Lambda}^{p_h}$), daily ($\boldsymbol{\Lambda}^{p_d}$) and weekly ($\boldsymbol{\Lambda}^{p_w}$) traffic transitional regularities. In particular, we estimate the importance score among granularity-specific embedding vectors by performing parametric matrix-based sum

operation as follows:

$$\mathbf{\Lambda} = \mathbf{W}^{p_h} \circ \mathbf{\Lambda}^{p_h} + \mathbf{W}^{p_d} \circ \mathbf{\Lambda}^{p_d} + \mathbf{W}^{p_w} \circ \mathbf{\Lambda}^{p_w} \tag{9}$$

where $\mathbf{W}^{p_h}$, $\mathbf{W}^{p_d}$ and $\mathbf{W}^{p_w}$ represents the learnable transformation matrices for different granularity-aware representations. With the element-wise multiplication $\circ$, we can generate the final pattern representation $\mathbf{\Lambda}$, with the explicitly exploration of spatial-temporal dependencies under the cross-granularity learning scenario.

3.4 External Factor Fusion

The traffic transitional regularities are also affected by various external factors, such as meteorological conditions and external temporal information. Hence, in the prediction scenario of traffic flow, it is also crucial to account for the influences of such external data sources which are defined as follows with details:

Definition 4. ***External Factors***. *We consider four types of external factors as the complementary data sources, namely, Weather conditions, Temperature/$^\circ$C, Wind speed/mph and Holiday signals. Each type of data source is associated with several encoding vectors with respect to different feature dimensions. Considering the weather condition as a concrete example, we generate four vectors ($f_{sun} \in \mathbb{R}^T$, $f_{rain} \in \mathbb{R}^T$, $f_{fog} \in \mathbb{R}^T$, $f_{snow} \in \mathbb{R}^T$) corresponding to sunny, rainy, foggy and snowy, respectively. Each element $f_*^t = 1$ if t-th day is positive for the weather condition feature (e.g., rainy day), and $f_*^t = 0$ otherwise. Similar strategy is applied for encoding holiday signals, i.e., weekday, weekend and national holiday. Furthermore, we apply normalization to scale quantitative temperature and wind speed values into the range of $[0, 1]$ for each target time granularity (e.g., hour or half a hour).*

Based on the aforementioned definitions, we could associate each region with the external feature vector f_*^T. We utilize a multi-layer perceptron architecture to map f_*^T into a latent space with a representation of $\mathbf{H}_{m,n} \in \mathbf{H}$, where $\mathbf{H} \in \mathbb{R}^{M \times N \times d}$. We further perform the concatenation of $\mathbf{H}$ and $\mathbf{\Lambda}$, and feed it into the prediction layer with a feed-forward network structure to forecast the future traffic.

3.5 The Learning Process of ST-CGA

3.5.1 **Optimized Objective.** . In this work, we aim to simultaneously predict the input and output traffic volume of each region across the entire city with the following defined loss function:

$$\begin{aligned}\mathcal{L} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \lambda[(\bar{x}^i_{m,n,t}) - (x^i_{m,n,t})]^2 \\ +(1-\lambda)[(\bar{x}^o_{m,n,t}) - (x^o_{m,n,t})]^2\end{aligned} \tag{10}$$

where λ balances the influence of input and output traffic flow. $\bar{x}^i_{m,n,t}$ and $\bar{x}^o_{m,n,t}$ represents the ground truth traffic volume of input and output flow, at the region $r_{m,n}$ and t-th time slot.

3.5.2 **Complexity Analysis of ST-CGA Framework.** We then analyze the time complexity of our ST-CGA framework. The model first takes linear time complexity for data preparation. The resulted $|P|$ granularity-specific tensors use $O(|P| \times T \times M \times N \times d)$ to calculate the query, key and value matrices, and use $O(|P| \times T^2 \times d)$ for weighted summations. In the latter region-wise graph attention module, ST-CGA takes $O(|P| \times M \times N \times d \times d')$ for the high-level feature representation, and takes $O(|P| \times M^2 \times N^2 \times d')$ for computing the weights and the attentive aggregation. For the spatial relation modeling between regions, ST-CGA first employs $O(|P| \times M \times N \times d \times d^2_{\text{conv}} \times C)$ computations for each convolution, where d_{conv} is the size of the convolutional kernel. In the later channel-aware recalibration network, $O(|P| \times M \times N \times C^2 \times d^2_{\text{FConv}})$ is required for the fully convolutional approach and $O(|P| \times M \times N \times C \times d_{\text{FCL}})$ is required for the fully connected scheme, where d_{FConv} is the kernel size for the convolutions and d_{FCL} is the dimensionality of the FCL hidden layer. The external factor fusion takes the $O(|f| \times d)$ complexity where $|f|$ denotes the dimensionality of the external features. Overall, the $O(|P| \times M^2 \times N^2 \times d')$ computations from the graph attention module clearly dominate the time complexity of ST-CGA in most real-world cases.

4 EVALUATION

In this section, we perform extensive experiments on different real-world datasets to evaluate the performance of our developed *ST-CGA* framework and compete with state-of-the-art prediction techniques. The performance improvement over baselines show the effectiveness of our method. In particular, this section aims to answer the following research questions from different aspects:

- **RQ1**: How does our proposed *ST-CGA* perform when competing with state-of-the-art traffic flow prediction techniques?
- **RQ2**: How does the designed key modules contribute the forecasting performance of *ST-CGA* framework?
- **RQ3**: How is the forecasting performance of *ST-CGA* approach *w.r.t* different period granularity configurations in cross-granularity pattern integration?
- **RQ4**: What is the impact of different encoder functions in our channel-aware recalibration network?
- **RQ5**: How do different hyperparameter settings impact the prediction accuracy of *ST-CGA* framework?
- **RQ6**: What explainable relational patterns does *ST-CGA* capture across geographical regions during the traffic prediction?

In the following subsections, we first introduce the experimental settings and then present the evaluation results corresponding to the above research questions.

4.1 Experimental Settings

4.1.1 **Data Description.** To evaluate the performance of our *ST-CGA* prediction framework, we perform experiments on four real-world traffic flow datasets from two cities (Beijing–BJ and New York City–NYC) which are described with details as follows:

BJ-Taxi. This data contains 34,000+ processed taxi trajectories from Beijing over four time periods (07/01/2013-10/30/2013; 03/01/2014-06/30/2014; 03/01/2015-06/30/2015; 11/01/2015-04/10/2016) [34]. The entire urban area is divided into disjoint 32×32 region grids, and each taxi trajectory is mapped into individual region. The input and output traffic flow are measured every half an hour.

NYC-Taxi. NYC-Taxi dataset collects 22,000,000+ taxi trajectory records of New York City span from 01/01/2015 to 03/01/2015 [28].

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

We partition NYC into 10×20 regions and the traffic flow is measured with the time interval of half an hour.

NYC-Bike-1. It records the trajectories from the bike system in NYC, with the format of (bike id, start/end timestamp, start/end station coordinates) [34]. Each day is divided into 24 time slots to estimate the traffic volume of each region from the 16×8 grid map.

NYC-Bike-2. This is another bike trajectories collected from New York City Bike system from 07/01/2016 to 08/29/2016 [28]. It contains 2,600,000+ bike trajectory records and utilizes the half an hour as the measurement time interval with the 10×20 grid map.

We summarize the data source statistics in Table 1.

4.2 Evaluation Protocols

4.2.1 **Data Preprocessing.** Based on the time intervals for data measurement of each dataset, we set the period granularity set $P \in \{30min, day, week\}$ for BJ-Taxi, NYC-Taxi and NYC-Bike-2 data, and $P \in \{hour, day, week\}$ for NYC-Bike-1 data. To be consistent with the settings in [28], we filter out the instances with the traffic volume less than 10 which is less important in both industry and academy [29]. Additionally, during the final traffic prediction phase, we use the tanh function to learn non-linear feature interactions and utilize minmax normalization to map the input data to $[-1, 1]$ for maintaining consistency between input and output. In the evaluation, the final output is re-scaled to be consistent of original data ranges before compared with ground truth.

4.2.2 **Evaluation Metrics.** We adopt two commonly used evaluation metrics *Root Mean Squared Error (RMSE)* and *Mean Absolute Percentage Error (MAPE)* [14, 18].

4.2.3 **Methods for Comparison.** We compared *ST-CGA* with the following state-of-the-art traffic prediction methods from various methodology research lines:

Conventional Time Series Prediction Approaches:

- **Auto-Regressive Integrated Moving Average (ARIMA)** [17]. It is a widely used time series analysis model which gauges the relation strength of one dependent variable relative to other varying variables from the time series data.
- **Support Vector Regression (SVR)** [1]. It has been used for time series prediction via mapping the input data into a high dimensional feature space with a nonlinear function.

Traditional Hybrid Learning Approach:

- **Fuzzy+NN** [19]: This hybrid model is composed of a fuzzy input/output filter and a stacked feed-forward network structure to predict the traffic volume information.

RNN-based Methods for Spatial-Temporal Forecasting:

- **Recurrent Neural Networks (RNN)**: This method models the temporal contextual information from the generated traffic series data with the recurrent neural network architecture.
- **LSTM Network (LSTM)**: This baseline proposes to use long short-term memory layer to predict the future traffic data using the forget gate to address the exploding gradients of RNN.

CNN-based Traffic Flow Prediction Framework:

- **Deep Spatial-Temporal Network (DeepST)** [35]. It develops the convolutional network to model the neighboring spatial correlations and a global neural encoder to map temporal features.
- **Deep Spatio-Temporal Residual Networks (ST-ResNet)** [34]. This method employs residual networks to model spatial dependencies between different regions in a city and aggregates the temporal properties with the convolution operations.
- **Multi-View Spatial-Temporal Network (DMVST-Net)** [29]. This solution integrates the local convolutional network and LSTM encoder to jointly consider spatial-temporal correlations. Furthermore, graph embedding technique is applied to generate region representations.

Attention-based Traffic Prediction Model:

- **Spatial-Temporal Dynamic Network (STDN)** [28]. A periodically shifted attention mechanism is combined with a convolutional-recurrent architecture, to capture spatial-temporal signals.

Traffic Forecasting Techniques with Graph Neural Networks:

- **Spatio-Temporal Graph Convolution Network (ST-GCN)** [30]. It learns spatial-temporal patterns simultaneously from the graph structure-based traffic series data, with the integration of graph and gated temporal convolution.
- **Spatio-Temporal Multi-Graph Convolution Network (ST-MGCN)** [3]. It proposes a multi-graph convolutional framework to capture the pair-wise correlation among regions.

Deep Hybrid Traffic Flow Predictive Models.

- **Urban Flow Magnifier (UrbanFM)** [14]. This hybrid approach proposes an inference network to extract traffic flow distributions and a fusion sub-network to consider external factors.
- **Spatial-Temporal Meta-Learning Model (ST-MetaNet)** [18]. It leverages the meta knowledge to model the spatial-temporal patterns with a seq2seq model, based on a designed recurrent graph attentive network structure.

4.2.4 **Parameter Settings.** We implement *ST-CGA* with Tensorflow and leverage Adam as the optimizer. The latent dimensionality d and encoded channel dimension size C in our channel-aware recalibration network are set as 64. We utilize 3 graph attention layers to capture the inter-region traffic dependencies. The filter size in our spatial relation encoding process is set as 3×3. The length of input traffic series for different resolutions (*i.e.*, hour–T_h, day–T_d and week–T_w) is chosen from $\{1, 2, 3, 4, 5, 6\}$, $\{1, 2, 3, 4, 5\}$, $\{1, 2, 3, 4, 5, 6\}$, respectively. During the external factor fusion phase, we utilize the multilayer perceptron of 3 layers structure. All the methods are trained from scratch without any pre-training on a

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

Category	Datasets	BJ-Taxi				NYC-Bike-1			
	Metrics	RMSE		MAPE (%)		RMSE		MAPE (%)	
	Methods	Inflow	Outflow	Inflow	Outflow	Inflow	Outflow	Inflow	Outflow
Conventional	ARIMA	22.01±0.08	23.96±0.07	30.91±0.02	32.21±0.03	9.25±0.02	11.70±0.02	35.79±0.01	36.4±0.03
	SVR	21.46±0.10	22.07±0.06	22.66±0.04	22.31±0.04	8.62±0.03	9.04±0.02	23.51±0.03	24.02±0.05
	Fuzzy+NN	22.36±0.30	23.02±0.33	22.71±0.32	22.83±0.35	8.51±0.29	9.15±0.31	23.96±0.28	24.48±0.31
RNN-based	RNN	27.14±0.64	27.93±0.53	24.15±0.49	24.79±0.51	8.98±0.55	9.21±0.43	28.19±0.47	28.51±0.53
	LSTM	27.03±0.57	27.52±0.55	23.63±0.47	24.13±0.43	8.61±0.49	9.07±0.47	27.44±0.51	28.06±0.48
CNN-based	DeepST	19.32±0.21	21.02±0.42	22.47±0.39	22.53±0.41	7.62±0.19	8.14±0.20	22.77±0.36	23.17±0.43
	ST-ResNet	17.02±0.20	22.32±0.24	23.54±0.40	23.77±0.43	6.24±0.23	6.57±0.19	23.86±0.48	24.76±0.41
	DMVST-Net	16.63±0.35	17.12±0.30	22.58±0.28	23.09±0.31	5.76±0.21	6.03±0.17	22.44±0.38	23.62±0.34
Attention-based	STDN	15.21±0.12	18.62±0.20	21.01±0.24	22.14±0.19	4.44±0.24	5.88±0.16	21.72±0.29	22.58±0.26
Deep Hybrid	UrbanFM	15.18±0.16	18.45±0.15	20.51±0.21	20.89±0.27	3.92±0.15	4.59±0.13	21.51±0.31	22.43±0.2
	ST-MetaNet	15.01±0.29	18.28±0.25	19.95±0.26	20.77±0.28	3.89±0.18	4.66±0.16	21.26±0.39	22.16±0.19
GNN-based	ST-GCN	15.12±0.19	18.33±0.17	19.94±0.20	20.75±0.22	3.79±0.11	4.71±0.12	21.10±0.21	21.99±0.17
	ST-MGCN	15.07±0.20	18.27±0.19	19.93±0.19	20.73±0.19	3.74±0.16	4.65±0.21	21.08±0.22	21.96±0.16
Our Model	***ST-CGA***	**14.59±0.21**	**17.55±0.17**	**19.02±0.18**	**20.28±0.18**	**3.02±0.05**	**4.00±0.07**	**20.62±0.19**	**21.45±0.17**
Category	Datasets	NYC-Taxi				NYC-Bike-2			
	Metrics	RMSE		MAPE (%)		RMSE		MAPE (%)	
	Methods	Inflow	Outflow	Inflow	Outflow	Inflow	Outflow	Inflow	Outflow
Conventional	ARIMA	27.25±0.02	36.53±0.03	20.91±0.06	22.21±0.04	11.25±0.02	11.53±0.01	25.79±0.07	26.53±0.06
	SVR	26.15±0.05	34.67±0.03	18.22±0.05	20.93±0.02	10.12±0.03	11.01±0.07	23.45±0.04	23.99±0.03
	Fuzzy+NN	26.01±0.34	34.54±0.31	18.97±0.27	21.55±0.29	11.33±0.30	11.86±0.28	24.69±0.27	25.13±0.34
RNN-based	RNN	30.18±0.56	38.33±0.55	26.24±0.49	26.89±0.51	13.55±0.48	15.37±0.47	27.03±0.43	27.61±0.50
	LSTM	29.99±0.47	37.52±0.49	25.93±0.50	26.48±0.49	13.67±0.48	15.90±0.50	27.18±0.47	27.89±0.46
CNN-based	DeepST	23.61±0.23	26.84±0.19	22.39±0.33	22.41±0.35	7.62±0.19	8.14±0.20	22.77±0.26	23.17±0.23
	ST-ResNet	21.76±0.24	26.35±0.32	21.16±0.46	21.28±0.50	8.87±0.16	9.84±0.16	23.01±0.42	23.11±0.37
	DMVST-Net	20.66±0.42	25.83±0.24	17.23±0.30	17.46±0.40	8.73±0.21	9.30±0.19	21.73±0.41	22.36±0.33
Attention-based	STDN	19.35±0.27	24.25±0.22	16.48±0.23	16.62±0.21	8.29±0.18	8.99±0.16	21.25±0.31	22.27±0.31
Deep Hybrid	UrbanFM	19.10±0.27	24.12±0.30	16.33±0.23	16.47±0.18	8.22±0.14	8.91±0.12	21.29±0.31	22.28±0.26
	ST-MetaNet	18.33±0.21	23.92±0.26	16.21±0.24	16.30±0.19	8.17±0.02	8.89±0.10	21.19±0.27	21.79±0.21
GNN-based	ST-GCN	17.98±0.14	23.04±0.22	15.91±0.23	15.92±0.18	8.03±0.19	8.73±0.12	21.22±0.20	21.94±0.19
	ST-MGCN	17.92±0.18	22.96±0.20	15.88±0.21	15.90±0.21	7.91±0.17	8.70±0.17	21.18±0.19	21.72±0.18
Our Model	***ST-CGA***	**17.11±0.19**	**22.12±0.16**	**15.20±0.15**	**15.34±0.18**	**7.34±0.15**	**8.42±0.10**	**20.66±0.17**	**21.02±0.17**

single NVIDIA GeForce GTX 1080 Ti GPU with a learning rate and batch size of $1e^{-3}$ and 32.

4.3 Performance Comparison (RQ1)

We evaluate the performance of our *ST-CGA* and compare it with state-of-the-art baselines and present the traffic prediction accuracy on four datasets in Table 2. We observe that *ST-CGA* consistently outperforms other baselines by obtaining high improvements over different types of state-of-the-art predictive models. We attribute such performance improvements to the learning of global inter-region traffic dependencies and channel-aware spatial relations, under the multi-resolution representation paradigm. Our developed *ST-CGA* integrative framework not only characterizes the multi-grained traffic variation patterns, but also learns inter-dependencies between both nearby and distant regions.

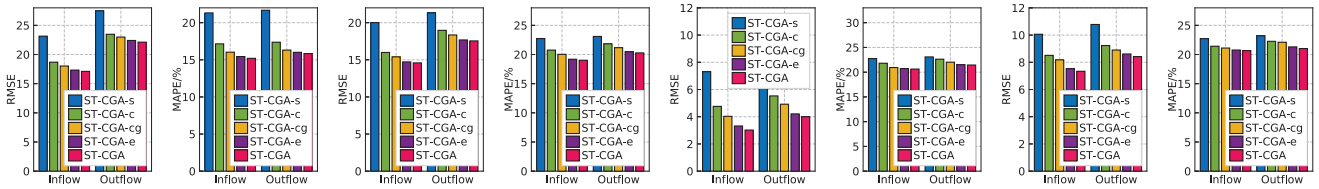
Additionally, we could notice that traffic predictive solutions with graph neural network to model the global region relations (*i.e.*, ST-GCN and ST-MGCN), achieves better performance than other baselines in most cases, which justifies the utility of capturing the spatial contexts from a global perspective. Among various competitive methods, the deep hybrid predictive models (*i.e.*, UrbanFM and ST-MetaNet) outperforms the mere utilization of i) recurrent architecture methods (RNN and LSTM), and ii) convolutional network approaches (DeepST and ST-ResNet). This performance gap indicates that only considering spatial or temporal dimensions can hardly handle the traffic evolving patterns with complex multi-resolution periodicity and arbitrary spatial dependencies.

4.4 Model Ablation Study (RQ2)

To get a better understanding of the designed integrative traffic prediction framework, we perform ablation experiments over several key modules of *ST-CGA* to investigate their component-wise impact. We introduce the following model variants corresponding to different sub-components.

- **Effect of Channel-Aware Recalibration Module**: *ST-CGA-c*. We do not include the recalibration network across position-aware channels, to generate the channel-specific importance weights with the learned mask tensor.
- **Effect of Graph Attention Module**: *ST-CGA-cg*. Another variant of *ST-CGA* which directly captures the spatial and temporal correlations with the integration of self-attention and convolution-based residual unit.
- **Effect of Region-wise Spatial Relation Modeling**: *ST-CGA-s*. This variant removes the spatial relation encoder across different regions, *i.e.*, do not utilize the integrative architecture of channel-aware convolutional network.
- **Effect of External Factor Modeling**: *ST-CGA-e*. We do not consider the external factors (*e.g.*, meteorological data) as additional data sources in *ST-CGA* to assist the traffic prediction task.

The evaluation results on both inflow and outflow traffic are reported in Figure 2. We can notice that our integrative framework *ST-CGA* outperforms the compared model variants in all cases, which can draw the following conclusions: (1) The performance gain between *ST-CGA* and *ST-CGA-cg* confirms the effectiveness of the graph attention module to tackle with the complex inter-region traffic dependencies; (2) Differentiating the importance of



(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-CGA framework in terms of RMSE and MAPE.

different channel-specific representations is helpful to capture spatial semantics from micro perspective; (3) While both spatial and temporal correlations are modeled with the self-attentive convolutional network, the ignorance of learning global traffic pattern inter-dependencies and channel-aware weight mask tensor will degrade the model performance; (4) The exploration of spatial relational structures is necessary to augment the modeling of high-level spatial contextual signals; (5) The external factors can assist the neural architecture for better traffic prediction, when they are property incorporated into our *ST-CGA*.

4.5 Effect of Multi-Resolution Dynamics (RQ3)

We investigate whether considering different configurations of granularity-specific temporal patterns, is beneficial to the traffic forecasting task. Towards this end, we examine *ST-CGA* with different settings of period granularity set P:

- $ST\text{-}CGA_h$: $P \in \{hour/30mins\}$
- $ST\text{-}CGA_{h,d}$: $P \in \{hour/30mins, day\}$
- $ST\text{-}CGA_{h,w}$: $P \in \{hour/30mins, week\}$
- $ST\text{-}CGA_{h,d,w}$: $P \in \{hour/30mins, day, week\}$

The evaluation results are reported in Figure 3. We can notice that $ST\text{-}CGA_{h,d,w}$ achieves the best performance as compared to other variants, which justifies that incorporating more time granularities into our channel-aware convolutional graph attention network is helpful for learning more accurate temporal patterns of traffic flow. Additionally, modeling the temporal dimension of traffic data from two granularities (*i.e.*, <hourly, daily> or <hourly, weekly>) could improve the performance, as compared with the temporal dependency learning with singular time granularity.

4.6 Impact of Encoder Functions in Channel-Aware Recalibration (RQ4)

We further investigate the impact of employing different encoder functions (fully convolutional network: $ST\text{-}CGA_{FConv}$; fully connected layers: $ST\text{-}CGA_{FCL}$) in our channel-aware recalibration network. In particular, the convolutional encoder is configured with a filter size of 3×3 and 64 kernels. The corresponding stride is set as (2,2). Moreover, the non-linear transformation function in the fully connected encoder aims to transform $(M \times N \times d)$ to $(M \times N \times C)$ from top-down to bottom-up architecture. As shown in Table 3, we can observe that $ST\text{-}CGA_{FCL}$ achieves better performance when competing with $ST\text{-}CGA_{FConv}$. The potential reason is that fully connected layers might supercharge the convolution-based residual unit with modeling of high-level non-linearities. $ST\text{-}CGA_{FConv}$ may provide redundant relation encoding effects as convolution-based residual unit (with similar convolution framework over $M \times N$ dimension), which can hardly comprehensively capture such non-linear cross-channel feature interaction.

Table 3: Effect Investigation of encoder functions in the channel-aware recalibration network on NYC-Taxi data.

Metrics	Inflow		Outflow	
Methods	RMSE	MAPE	RMSE	MAPE
$ST\text{-}CGA_{FConv}$	17.83±0.22	15.72±0.16%	22.91±0.28	16.03±0.15%
$ST\text{-}CGA_{FCL}$	**17.11±0.19**	**15.20±0.15%**	**22.12±0.16**	**15.34±0.18%**

4.7 Hyperparameter Study (RQ5)

Next, we investigate the influence of key hyperparameters for our *ST-CGA* framework. Figure 4 shows the prediction accuracy on NYC-Taxi data under different settings of parameters. When varying the target parameter, we set other parameters as default values. We summarize the key observations as:

Filter Size. We observe that filter size with a larger spatial coverage does not necessarily result in better traffic prediction performance. One potental reason is that when performing convolution operation with larger kernel size, more hyperparameters need to be learned, which may increase the training difficultly of *ST-CGA*.

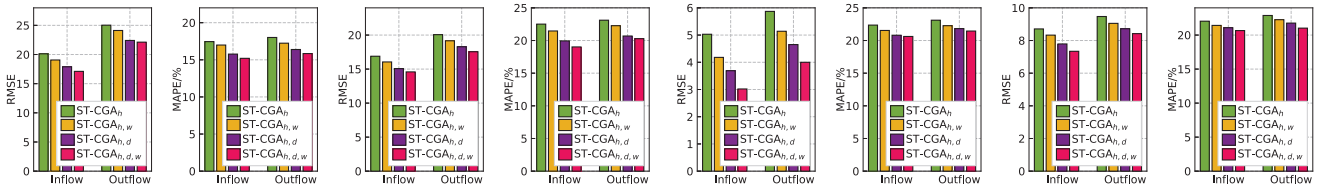
Encoded Channel Dimensionality. We notice that the increase of encoded channel dimensions benefits the prediction at the early stage ($C = 64$), but the performance degrades with larger channel dimension size. This may be caused by the overfitting phenomenon.

Resolution-specific Sequence Length. Our *ST-CGA* model could achieve comparable forecasting accuracy only when $T_h = 4$ and $T_d = 3$, which justifies the effectiveness of *ST-CGA* in capturing the long-term temporal dynamics without involving long traffic series.

Number of Graph Attention Layers. Increasing the depth of graph attention framework could enhance the traffic prediction results. Obviously, *ST-CGA* with 2 and 3 attention layers outperforms the model which considers first-order neighbors only. We attribute such improvement to the effective encoding of inter-region traffic dependencies, with high-order connectivities. However, when we further stack 4 and 5 graph-based propagation layers, the overfitting can be observed and may stem from the introduction of noise into the pattern representation.

4.8 Case Study: Model Interpretation (RQ6)

We conduct case studies to investigate the explainability of our *ST-CGA*, by visualizing the attentive weights of graph inter-region dependency encoder (as shown in Figure 5). Given the target region: "Dongzhimen Bridge" in Beijing, we find that its highly relevant regions which are either spatially neighbors or share similar region



(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 3: Effect of multi-resolution dynamics in terms of RMSE and MAPE.

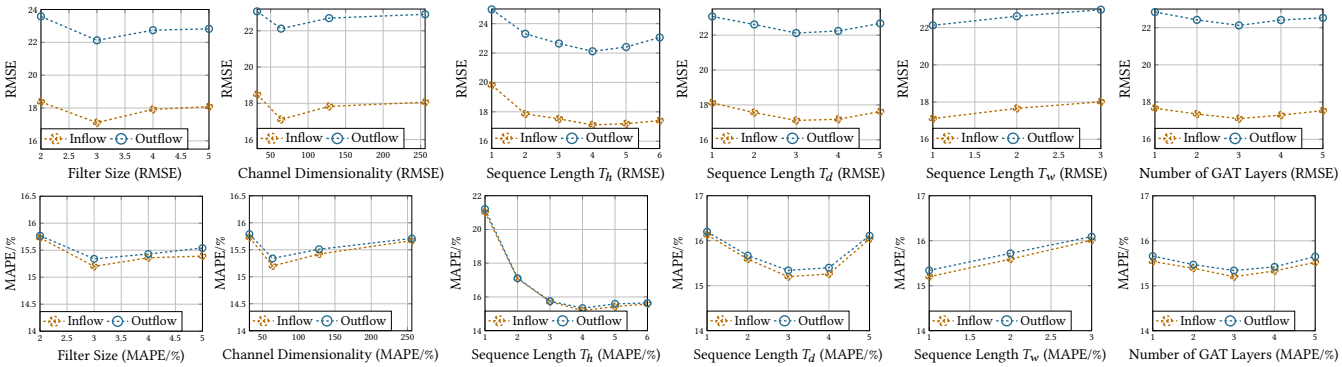


Figure 4: Hyper-parameter study on NYC-Taxi data in terms of RMSE and MAPE.

functionalities (*e.g.*, transportation hub and shopping center). Particularly, we extract several qualitative examples to better understand the learned spatial dependency representations across regions in urban space, which further verifies the rationality of *ST-CGA* in capturing both local and global cross-region geographical relations.

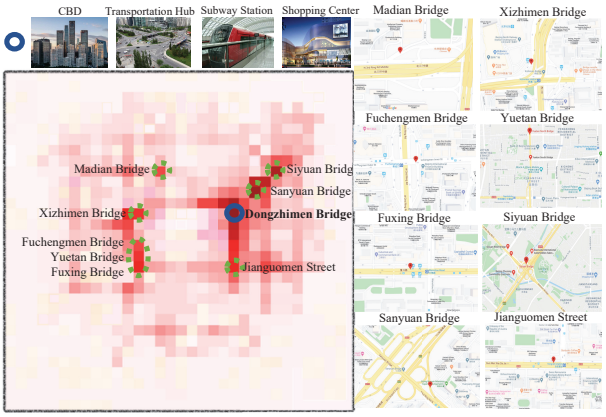


Figure 5: Case study from BJ-Taxi data. The target region (Dongzhimen Bridge) has four key urban functions (*i.e.*, CBD, Transportation Hub, Subway Station, Shopping Center). Eight highly relevant regions sharing similar functions are highlighted.

5 RELATED WORK

Data-Driven Urban Sensing Tasks. Urban sensing has emerged as a new sensing paradigm that empowers average people to contribute their observations and measurements about the physical world. Numerous novel applications have been developed to address various challenges in smart cities applications [4, 11, 33]. More recent works have focused on addressing important challenges such as intelligent transportation [2, 32], privacy-preserving [27], public safety [10], business assessment [13] and human activity modeling [7]. In particular, Chen *et al.* developed a dynamic cluster-based framework to predict over-demand occurrence in bike sharing systems [2]. Yang *et al.* addressed the privacy leakage problem in location-based services by proposing to obfuscate check-in data such that the privacy leakage is minimized under a given data distortion budget [27]. However, the traffic flow prediction problem in urban sensing still remains as a challenging problem to be solved. In this paper, we develop an end-to-end model to tackle this challenge.

Neural Network-based Traffic Flow Prediction. There exists a rich amount of work on the topic of traffic prediction using various deep learning techniques. For example, recurrent neural networks have been employed to explore temporal correlations of time-ordered traffic data [16, 31]. In addition, many efforts have been devoted to developing convolutional architecture to capture spatial dependencies with the image-based traffic data [29, 34]. Several hybrid traffic predictive solutions were introduced with meta-learning framework (ST-MetaNet) [18] and data fusion network (UrbanFM) [14]. Different from those approaches, this work collectively explore inherent spatial, temporal and semantic relations across regions, time periods as well as latent representation spaces, from both local to global perspectives.

Graph Neural Networks. Recently, graph neural networks (GNN) have been demonstrated to be effective in learning on graph structure data, such as user-item interaction graph [22] and item-item relation graph [8]. Motivated by GNN-based propagation models,

the graph convolutional neural network has been applied to the traffic prediction scenario to model the spatial region correlations [30]. Geng *et al.* [3] augment the graph convolutional architecture in demand forecasting with a multi-dimensional spatial-temporal graph. In this work, the designed ST-CGA is built upon the graph attention network and further incorporates multi-granularity relation structures into the spatial-temporal graph learning paradigm.

6 CONCLUSION

This work investigated the traffic flow prediction problem by developing a ST-CGA framework that collectively i) captures the multiple granularity-aware temporal factors that govern the dynamic transition regularities of traffic flow; ii) models the high-order spatial relation structures with a channel-aware convolutional graph learning model; iii) integrates the collaborative signals from spatial, temporal and semantic dimensions. We conduct extensive experiments on four real-world datasets. The results demonstrate the superiority of ST-CGA over many strong baselines. One possible direction for future work is to incorporate more external data sources (*e.g.*, region's function description) into our traffic prediction framework, to provide external signals for characterizing the correlations between regions.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their constructive feedback. This work was supported by National Nature Science Foundation of China (61672241), Major Project of National Social Science Foundation of China (18ZDA062), Natural Science Foundation of Guangdong Province (2016A030308013), Science and Technology Program of Guangdong Province (2019A050510010), and Fundamental Research Funds for the Central Universities (x2js-D2192830).

REFERENCES

[1] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *Transactions on Intelligent Systems and Technology (TIST)* 2, 3 (2011), 27.

[2] Longbiao Chen, Daqing Zhang, et al. 2016. Dynamic cluster-based over-demand prediction in bike sharing systems. In *International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. ACM, 841–852.

[3] Xu Geng, Yaguang Li, Leye Wang, Lingyu Zhang, Qiang Yang, Jieping Ye, and Yan Liu. 2019. Spatiotemporal Multi-Graph Convolution Network for Ride-hailing Demand Forecasting. In *International Conference on Artificial Intelligence (AAAI)*.

[4] Daniel Gooch, Annika Wolff, Gerd Kortuem, and Rebecca Brown. 2015. Reimagining the role of citizens in smart city projects. In *International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. ACM, 1587–1594.

[5] Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. 2019. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *International Conference on Artificial Intelligence (AAAI)*, Vol. 33. 922–929.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[7] Chao Huang, Dong Wang, and Shenglong Zhu. 2017. Where are you from: Home location profiling of crowd sensors from noisy and sparse crowdsourcing data. In *International Conference on Computer Communications (Infocom)*. IEEE, 1–9.

[8] Chao Huang, Xian Wu, Xuchao Zhang, Chuxu Zhang, Jiashu Zhao, Dawei Yin, and Nitesh V Chawla. 2019. Online purchase prediction via multi-scale modeling of behavior dynamics. In *International Conference on Knowledge Discovery and Data Mining (KDD)*. 2613–2622.

[9] Chao Huang, Chuxu Zhang, Jiashu Zhao, Xian Wu, Dawei Yin, and Nitesh Chawla. 2019. Mist: A multiview and multimodal spatial-temporal learning framework for citywide abnormal event forecasting. In *The Web Conference (WWW)*. 717–728.

[10] Chao Huang, Junbo Zhang, Yu Zheng, and Nitesh V Chawla. 2018. DeepCrime: attentive hierarchical recurrent networks for crime prediction. In *International Conference on Information and Knowledge Management (CIKM)*. 1423–1432.

[11] Li Jin, Zhuonan Feng, and Ling Feng. 2016. A Context-aware Collaborative Filtering Approach for Urban Black Holes Detection. In *International Conference on Information and Knowledge Management (CIKM)*. ACM, 2137–2142.

[12] Jiacheng Li, Yujie Wang, and Julian McAuley. 2020. Time Interval Aware Self-Attention for Sequential Recommendation. In *International Conference on Web Search and Data Mining (WSDM)*. 322–330.

[13] Jianxun Lian, Fuzheng Zhang, Xing Xie, and Guangzhong Sun. 2017. Restaurant survival analysis with heterogeneous information. In *The Web Conference (WWW)*. ACM, 993–1002.

[14] Yuxuan Liang, Kun Ouyang, Lin Jing, Sijie Ruan, Ye Liu, Junbo Zhang, et al. 2019. UrbanFM: Inferring Fine-Grained Urban Flows. In *International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 3132–3142.

[15] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, et al. 2017. Feature pyramid networks for object detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2117–2125.

[16] Qiang Liu, Shu Wu, Liang Wang, et al. 2016. Predicting the Next Location: A Recurrent Model with Spatial and Temporal Contexts.. In *International Conference on Artificial Intelligence (AAAI)*. 194–200.

[17] Bei Pan, Ugur Demiryurek, et al. 2012. Utilizing real-world transportation data for accurate traffic prediction. In *ICDM*. IEEE, 595–604.

[18] Zheyi Pan, Yuxuan Liang, Weifeng Wang, et al. 2019. Urban Traffic Prediction from Spatio-Temporal Data Using Deep Meta Learning. In *International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM.

[19] Dipti Srinivasan, Chee Wai Chan, and PG Balaji. 2009. Computational intelligence-based congestion prediction for a dynamic urban street network. *Neurocomputing* 72, 10-12 (2009), 2710–2716.

[20] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. In *ICLR*.

[21] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, et al. 2017. Residual attention network for image classification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 3156–3164.

[22] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *International Conference on Research and Development in Information Retrieval (SIGIR)*. 165–174.

[23] Xian Wu, Baoxu Shi, Yuxiao Dong, Chao Huang, Louis Faust, and Nitesh V Chawla. 2018. Restful: Resolution-aware forecasting of behavioral time series data. In *International Conference on Information and Knowledge Management (CIKM)*. 1073–1082.

[24] Lianghao Xia, Chao Huang, Yong Xu, Peng Dai, Bo Zhang, and Liefeng Bo. 2020. Multiplex Behavioral Relation Learning for Recommendation via Memory Augmented Transformer Network. In *International Conference on Research and Development in Information Retrieval (SIGIR)*. 2397–2406.

[25] Mingzhou Xu, Derek F Wong, Baosong Yang, Yue Zhang, and Lidia S Chao. 2019. Leveraging local and global patterns for self-attention networks. In *Annual Meeting of the Association for Computational Linguistics (ACL)*. 3069–3075.

[26] Baosong Yang, Jian Li, Derek F Wong, Lidia S Chao, Xing Wang, and Zhaopeng Tu. 2019. Context-aware self-attention networks. In *International Conference on Artificial Intelligence (AAAI)*, Vol. 33. 387–394.

[27] Dingqi Yang, Daqing Zhang, Bingqing Qu, and Philippe Cudré-Mauroux. 2016. PrivCheck: privacy-preserving check-in data publishing for personalized location based services. In *International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. ACM, 545–556.

[28] Huaxiu Yao, Xianfeng Tang, Hua Wei, et al. 2019. Revisiting Spatial-Temporal Similarity: A Deep Learning Framework for Traffic Prediction. In *International Conference on Artificial Intelligence (AAAI)*.

[29] Huaxiu Yao, Fei Wu, Jintao Ke, Xianfeng Tang, et al. 2018. Deep multi-view spatial-temporal network for taxi demand prediction. In *International Conference on Artificial Intelligence (AAAI)*. 2588–2595.

[30] Bing Yu, Haoteng Yin, and Zhanxing Zhu. 2018. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *International Joint Conferences on Artificial Intelligence (IJCAI)*.

[31] Rose Yu, Yaguang Li, Cyrus Shahabi, et al. 2017. Deep learning: A generic approach for extreme condition traffic forecasting. In *SIAM International Conference on Data Mining (SDM)*. SIAM, 777–785.

[32] Zhuoning Yuan, Xun Zhou, and Tianbao Yang. 2018. Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In *International Conference on Knowledge Discovery & Data Mining (KDD)*. 984–992.

[33] Fuzheng Zhang, Nicholas Jing Yuan, et al. 2015. Sensing the pulse of urban refueling behavior: A perspective from taxi mobility. *Transactions on Intelligent Systems and Technology (TIST)* 6, 3 (2015), 37.

[34] Junbo Zhang, Yu Zheng, and Dekang Qi. 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *International Conference on Artificial Intelligence (AAAI)*.

[35] Junbo Zhang, Yu Zheng, Dekang Qi, Ruiyuan Li, and Xiuwen Yi. 2016. DNN-based prediction model for spatio-temporal data. In *International Conference on Advances in Geographic Information Systems (SIGSPATIAL)*. 1–4.