Traffic Accident Risk Prediction via Multi-View Multi-Task Spatio-Temporal Networks

Senzhang Wang, *Member, IEEE,* Jiaqiang Zhang, Jiyue Li, Hao Miao, and Jiannong Cao, *Fellow, IEEE*

Abstract—Abnormal traffic incidents such as traffic accidents have become a significant health and development threat with the rapid urbanization of many countries. Thus it is critically important to accurately forecast the traffic accident risks of different areas in a city, which has attracted increasing research interest in the research area of urban computing. The challenges of accurate traffic risk forecasting are three-fold. First, traffic accident data in some areas of a city is sparse, especially for a fine-grained prediction, which may cause the zero inflation problem during model training. Second, the spatio-temporal correlations of the traffic accidents occurring in different areas are rather complex and non-linear, which is difficult to capture by existing shallow models like regression. Third, the occurrence of traffic accidents can be significantly affected by various context features including weather, POI and road network features. It is non-trivial to capture the complex associations between the diverse context features and traffic accident risks for building an accurate prediction model. To address the above challenges, this paper proposes a Multi-View Multi-Task Spatio-Temporal Networks (MVMT-STN) model to forecast fine- and coarse-grained traffic accident risks of a city simultaneously. Specifically, to address the data sparsity issue in a fine-grained prediction, we adopt a multi-task learning framework to jointly forecast both fine- and coarse-grained traffic accident risks by considering their spatial associations. For each granularity prediction, we design the channel-wise CNN and multi-view GCN to capture the local geographic dependency and global semantic dependency, respectively. In order to obtain the diverse impacts of the context features on traffic accidents, we also introduce a fusion learning module that integrates the channel-wise and multi-view features learned from different types of the external factors. We conduct extensive experiments over two large real traffic accident datasets. The results show that MVMT-STN improves the performance of traffic accident risk prediction in both fine- and coarse-grained prediction by a large margin compared with existing state-of-the-art methods.

Index Terms—Traffic accident forecasting, Multi-task learning, Spatio-temporal data, Graph Neural Networks

1 Introduction

With the fast urbanization, the number of vehicles in many big cities expands rapidly in the past several decades, which leads to the significant increase of traffic accidents. Traffic accidents have become one of the most significant public safety issues for many countries. According to the statistics of the World Health Organization (WHO), traffic accidents cause about 1.3 million deaths and 20 to 50 million non-fatal injuries in the world each year [1]. Therefore, accurate traffic accident forecasting is becoming crucially important to help the government and policymakers adopt certain traffic control or strategies to reduce the harm caused by traffic accidents [2]. For drivers, knowing the high risk areas of traffic accidents they are going to pass by in advance can also alert them to drive carefully for avoiding accidents.

Traditionally, statistical and linear regression based methods such as SVM and ARIMA [3] are widely used to predict traffic accidents by considering the number of traffic accidents of an area or road segment as time series data. The major limitation of such methods is that the complex spatio-temporal correlations of traffic accidents among different areas cannot be effectively captured. With the recent advances of deep learning techniques, various deep learning models such as LSTM [4], CNN [5] and autoencoder (AE) [6] are applied to forecast the cite-wide traffic accidents. Deep learning models are more effective to learn the non-linear spatio-temporal correlations of the traffic accident data, and thus can usually achieve better prediction performance. [7] proposed a Hetero-Convolutional Long Short-Term Memory (Hetero-ConvLSTM) model to capture spatial heterogeneity and temporal auto-correlation. But the influence of external context features such as POIs and road features are not fully considered. [8] designed a deep Dynamic Fusion Network framework to learn the complex time correlation and external factors' influence on traffic accidents. However, it only considers the local spatial correlations, but ignores the global semantic correlations. Although plentiful works have been conducted on traffic accident forecasting, we argue that existing methods still cannot yield satisfactory forecasting performance due to the following challenges.

First, the traffic accident data can be very sparse in some areas of a city, especially in the case that a fine-grained prediction (e.g. predicting the traffic accidents in small areas) is needed. As the typical abnormal traffic incidents, traffic accidents are more likely to occur in only a small number of areas of a city in some particular time intervals, while for most other areas there are no traffic accidents at all. For example, as shown in Fig.1, there are only four traffic

- *S.Z. Wang (szwang@csu.edu.cn) is with the School of Computer Science and Engineering, Central South University, Changsha, China, 410083.*
- *J.Q. Zhang and J.Y. Li are with College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China, 211106.*
- *H. Miao is with Department of Computer Science, Aalborg University, Danmark.*
- *J.N. Cao is with Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China.*

Manuscript received April 19, 2005; revised August 26, 2015.

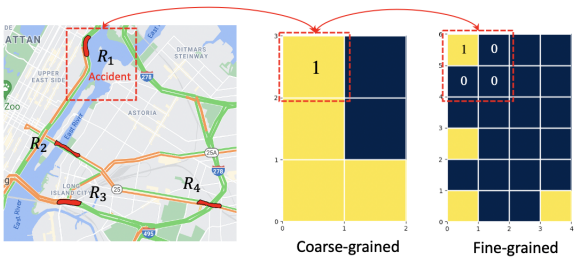

Fig. 1: Illustration of the data sparsity issue

accidents occurred in the selected 1-hour interval in regions $R1$, $R2$, $R3$, $R4$ of New York City, which is sparse. The data becomes even sparser if we partition the area into fine-grained cell regions as shown in the right part of the figure. Only one region marked in yellow in the red square has an traffic accident, while the number is 0 for all the other cell regions. The data sparsity issue will lead to the zero inflation issue for the training of deep learning models [9], making the models difficult to train.

Second, the spatio-temporal correlations of the traffic accident data are complex and non-linear, which are hard to capture by existing methods. Current deep learning methods generally model the traffic accident data from the temporal and spatial feature views separately. For the spatial feature view, [5, 10] divided a city into equal sized regions based on latitude and longitude, and then used convolution neural networks (CNN) for spatial feature learning. [11, 12] modeled the regions of city as a graph based on their spatial connectivity, and then graph convolution networks (GCN) were used to model the spatial correlations. A major limitation of the CNN and GCN based methods for spatial feature learning is that only the local spatial proximity is captured, while the global semantic correlations are largely ignored. Two geographically distant areas may present similar pattern of traffic accidents because they have similar POI distributions, road features and functionalities (e.g. business districts or residential districts) [13]. Therefore, it is necessary to capture the global semantic and local geographic correlations at the same time for traffic accident prediction. For the temporal feature view, RNN is usually used to learn the temporal correlations of the traffic accident time series data [4]. In traffic accident prediction, however, the temporal dependency is complex including both short- and long-term temporal correlations. The complex temporal auto-correlation is also hard to capture by existing RNN based models. Thus we need to automatically find the more relevant data while ignore the less relevant ones from historical data to improve the prediction performance.

Third, it is non-trivial to model the diverse effects of external context features on traffic accidents. Previous studies show that the causes of traffic accidents are complex and may be significantly affected by various external factors including weather, road features and POI [2, 14, 15]. It is also showed that human mobility data are highly relevant to the occurrence of traffic accidents [16]. Meanwhile, the impacts of different context features on traffic accidents over different areas can be diverse and change over time. For example, in a heavy rainy day, the context feature of weather will have a larger impact on abnormal traffic incidents compared with other factors. While in normal days, traffic accidents may be more correlated to POIs and road features. Existing work mostly directly concatenate the external features without considering their diverse impacts on traffic accidents [6, 7]. A more sophisticated model which can capture the diverse impacts of context features is required.

To address the above challenges, this paper proposes a Multi-View Multi-Task Spatio-Temporal Networks (MVMT-STN) model to forecast fine- and coarse-grained traffic accident risks of a city simultaneously. Specifically, we divide the city into cell regions of different granularities based on the latitude and longitude. For each granularity, we first propose a channel-wise CNN on the traffic accident risk spatial maps to learn the local spatial feature representation, and then build multi-view graphs based on the similarity of context features (i.e. POI, road features, accident risk) among the cell regions. A multi-view self-attention GCN is proposed to learn the global semantic feature representations over the multi-view graphs. The two types of feature representations are next integrated with a designed feature fusion gate. Meanwhile, in order to model the diverse effects of different external factors, in channel-wise CNN, we propose to use global average pooling to learn the influence weights for different feature channels including weather, POI, human mobility and traffic accident. To capture the non-linear temporal dependency of the traffic accidents, an attention LSTM is also employed for temporal feature learning. Finally, we adopt a multi-task learning framework to jointly forecast both fine- and coarse-grained traffic accident risks by considering their spatial associations and addressing the data sparsity issue. In summary, the contributions of this work are as follows.

•We for the first time propose a multi-task learning framework to simultaneously forecast the fine- and coarse-grained traffic accident risks. A cross-scale feature fusion mechanism is designed to integrate the latent features of the two scale data. A structure constraint is also introduced between fine- and coarse-grained predictions to effectively address the data sparsity issue in fine-grained prediction.

• We propose a channel-wise CNN and multi-view GCN to capture both the local geographic and global semantic dependencies, as well as model the effect of the diverse external context features. We also introduce the Attention-LSTM module to capture the complicated dependency of traffic accidents in the time dimension.

• We propose a feature fusion component to integrate the channel-wise and multi-view feature representations. Extensive experiments are conducted to evaluate our model over two real-world datasets. The experiment results verify the superior performance of our proposal compared with state-of-the-art models.

The remainder of the paper is organized as follows. We will first introduce the preliminaries and give a formal problem definition in Section 2. Data preparation and the MVMT-STN model will be elaborated in Section 3. The multi-task learning objective function will be presented in Section 4. Section 5 will introduce our evaluation and results discussion. Section 6 will review the related work, followed by our conclusion on this work in Section 7.

2 PRELIMINARIES AND PROBLEM DEFINITION

Definition 1. **Cell region.** We partition a city R under study into a grid map with the size $H \times W$ based on the latitude and longitude. We denote all the cell regions of the grid map as $R = \{r_{1,1}, ..., r_{i,j}, ...r_{H,W}\}$. A coarse-grained partition of R contains $\frac{H}{k} \times \frac{W}{k}$ grid cells, where $k \in \mathbf{N}_+$ is the coarsening coefficient.

Definition 2. **Human mobility tensor.** Given the human mobility data (*e.g.* taxi trajectory) of a city, we denote the human mobility in all the cell regions in time slot t as a tensor $\mathcal{S}^t \in \mathbb{R}^{2\times H\times W}$, where 2 is the number of human mobility measures including inflows and outflows. Each entry $S^t_{1,h,w}$ denotes the inflow into cell region $r_{h,w}$, and entry $S^t_{2,h,w}$ denotes the outflow.

The distribution and density of POI in each cell region indicate the land function of the region [13] , so it is helpful for the prediction of traffic accident risk. Meanwhile, the road features (e.g. the length and width of the roads, the road type, etc.) are also highly correlated to the traffic conditions of a cell region [17]. Thus we also consider the land features defined as follows.

Definition 3. **Land features.** We collect the land features including POI and road features in each cell region, and denote POI and road features in all the cell region as $P \in \mathbb{R}^{d_p\times H\times W}$ and $F \in \mathbb{R}^{d_f\times H\times W}$, respectively, where d_p, d_f are the feature sizes.

Definition 4. **External features.** Traffic abnormal incidents are also highly correlated to the external features including weather, holiday and time of a day. We denote the external features of cell regions R as $E \in \mathbb{R}^{d\times H\times W}$, where d is the feature size.

Definition 5. **Traffic accident risk map.** Following the work [18], traffic accident risks can be categorized into the following three types: minor accident risk, injury accident risk and fatal accident risk. The three types of risks are assigned to the risk values 1, 2 and 3, respectively. The risk map in a time slot t can be denoted as a matrix $X^t \in \mathbb{R}^{H\times W}$, whose each entry $X^t_{i,j}$ denotes the risk value of region $r_{i,j}$ in time slot t. All the traffic accident risk maps in T time slots form a risk tensor $\mathcal{X} \in \mathbb{R}^{T\times H\times W}$.

Definition 6. **Multi-view spatial graph.** To fully capture the effect of external context features to the global semantic dependency, we model a city with cell regions R as a multi-view spatial graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$. Specially, $\mathcal{G}$ contains three views, the POI similarity graph view $G^p(\mathcal{V}, E^p)$, the road feature similarity graph view $G^f(\mathcal{V}, E^f)$ and the traffic accident risk similarity graph view $G^p(\mathcal{V}, E^r)$. Note that the three views share the same nodes but different edges. For each view, the nodes correspond to the cell regions, and the view feature (i.e. POI, road features and accident risk) similarities between the regions denote the weights of the edges. For each node (cell region), we select Top-K nodes with the largest feature similarity as its neighbors.

Based on the above terminology definitions, we formally define the studied problem as follows.

Problem Definition 1. Given the cell regions of a city R, the coarsening coefficient k, the historical traffic accident risk tensor $\mathcal{X}$, the human mobility tensors $\{\mathcal{S}^1, ...\mathcal{S}^t, ...\mathcal{S}^T\}$, the land features $\{P, F\}$, the external features E, and the multi-view spatial graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ in the previous T time slots, we aim to predict the fine- and coarse-grained traffic accident risk maps $X^{f_{T+1}}$, $X^{c_{T+1}}$ in the next time slot $T + 1$ simultaneously.

3 METHODOLOGY

Fig. 2 illustrates the framework of the proposed model, which contains three major steps, spatial feature learning step, temporal feature learning step and cross-scale multi-task prediction step. In the spatial feature learning step, to capture the complex spatial correlations, we extract the local spatial features through a channel-wise CNN module and the global semantic features through a multi-view GCN module separately. To address the data sparsity issue, we conduct the spatial feature learning for fine- and coarse-grained traffic accident risk data simultaneously. The channel-wise CNN is conducted on the traffic accident risk spatial maps in Definition 5, and the multi-view GCN is conducted over the multi-view spatial graphs in Definition 6. To fuse the spatial features of the two scale data to address the data sparsity issue, a cross-scale GCN is proposed. In the temporal feature learning step, an Attention LSTM module is used to learn the temporal dependency of the input data. To integrate the local geographic and global semantic features, a fusion gate is designed as shown in the right part of the figure. Finally, in the prediction step, a multi-task learning framework is adopted to couple fine- and coarse-grained prediction. Considering the spatial associations between the two data of two scales, a structure constraint loss is introduced in the final objective function. Next, we will introduce the model in detail.

3.1 Data Preparation

Based on Definition 5, we assign the values 1, 2, and 3 to the three types of traffic accident risks, respectively. To prepare the data for the proposed model, we first categorize the types of traffic accidents and count the corresponding accident numbers for each cell region. The traffic accident risk value for cell region $r_{i,j}$ in time slot t can be calculated as follows:

$$risk^t_{r_{i,j}} = \sum_{l=1}^{3} l * sum^t_{r_{i,j}}(l) \tag{1}$$

where l indicates the type of traffic accident, $sum^t_{r_{i,j}}(l)$ is the number of traffic accident of type l in the time slot t in cell region $r_{i,j}$.

The occurrences of traffic accidents are affected by various external context features. The right of Fig.3 shows the traffic accident number under the different weather conditions including clear, cloudy, rain, snow and mist in Chicago. The left of Fig. 3 shows the number of traffic accidents occurring in different POIs. One can see that the accident number varies significantly under different weather conditions and POIs, which means that weather and POI can have a significant impact on traffic accidents.

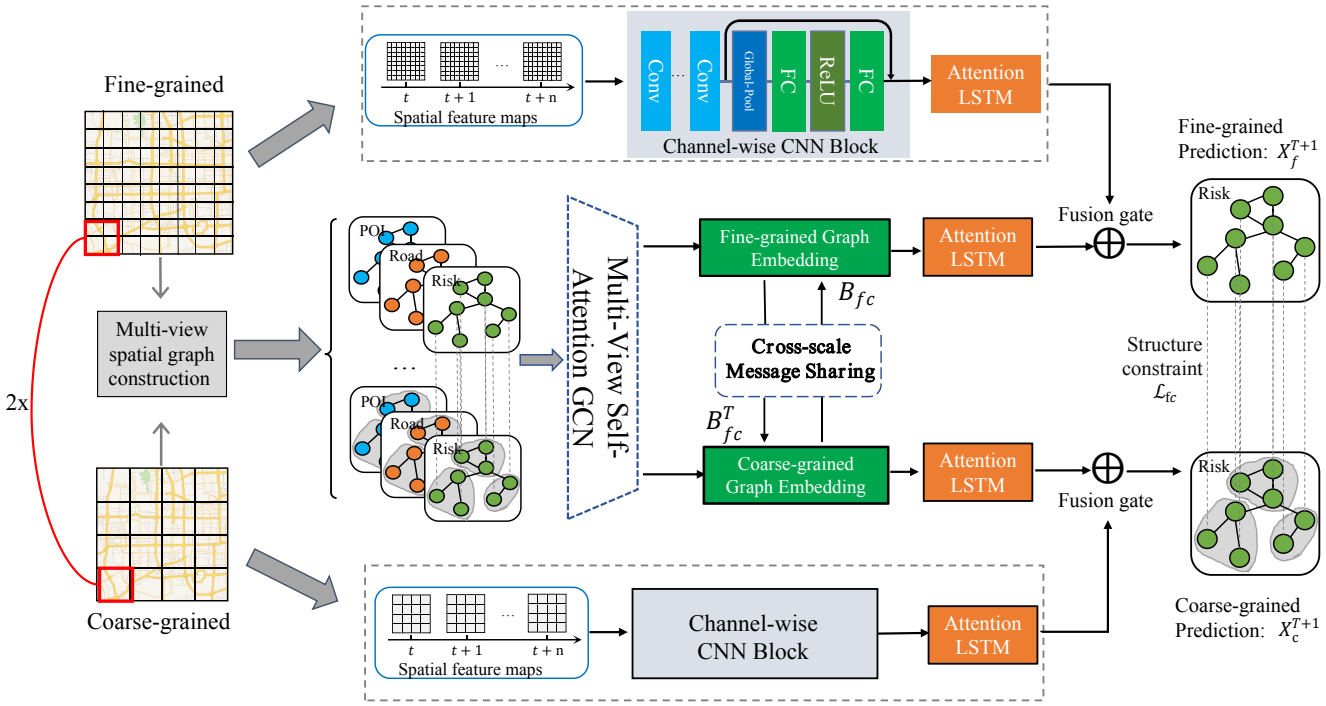

Fig. 2: The framework of the proposed MVMT-STN model.

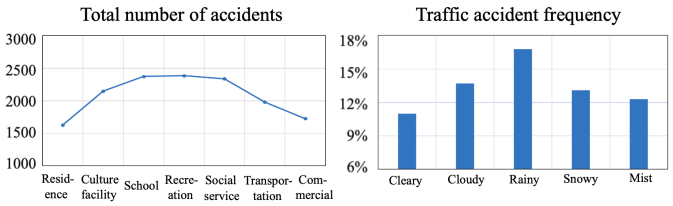

Fig. 3: The effect of weather conditions and POI categories to traffic accidents

Therefore, it is necessary to consider these external features to assist forecasting.

Previous studies [12, 19] showed that the current traffic accident risk of each cell region is highly correlated to the risks at previous time slots and the risks at the same time slot of previous days. To fully capture the short- and long-term temporal correlations, we use the traffic accident risk spatial maps in the previous κ time intervals as well as the maps at the same time slot in the previous ρ weeks as the input T time intervals ($T = \kappa + \rho$). Hence, the input data can be formulated as a union of two sets of historical traffic accident risk spatial maps as follows.

$$X_{input} = X_{short} \cup X_{long} = \{X_t \in R^T | t \in (1, 2, ..., \kappa) \cup t \in (1, 2, ..., \rho)\}) \quad (2)$$

3.2 Channel-Wise CNN Block for Local Spatial Feature Learning

The traffic conditions in geographically close cell regions are usually highly correlated, and Convolution Neural Networks (CNN) are widely used to capture the local spatial correlations among neighborhood regions for traffic accident risk prediction by previous works [10]. However, a major limitation of the general CNN model is that it only considers the traffic accident data in the local spatial dimension, but overlooks the impacts of various relevant features, including human mobility patterns and external features on the occurrence of traffic accidents. Therefore, we propose a channel-wise CNN block to capture the diverse impacts of land features and external features. Note that different from traditional multi-channel CNN models, we firstly calculate the attention coefficient on each channel via a global average pooling layer followed by two FC layers. The global average pooling layer is used to learn the high-level representation of each channel, and two FC layers are used to learn the nonlinear correlations for the representations. Then the original output is readjusted through the weights. For example, weather features may have a larger impact than static road features on traffic accident risks. Therefore, the channel-wise CNN should be able to learn the impact weights of different types of features on the occurrence of traffic accident.

Fig.4 shows the structure of the proposed channel-wise CNN block. It mainly consists of the multi-channel CNN and Squeeze-and-Excitation Networks (SENet) [20] to learn the local spatial correlations and capture the impacts of different features on traffic accident risks. Multi-channel CNN can learn and aggregate the representations of different feature channels on the local receptive field. The importance of different feature channels can be automatically learned by SENet. The SENet can selectively enhance the useful feature channels and suppress the useless feature channels, so as to realize the calibration of each feature channel. The channel-wise CNN block takes the multi-channel features including human mobility features, the land features, the

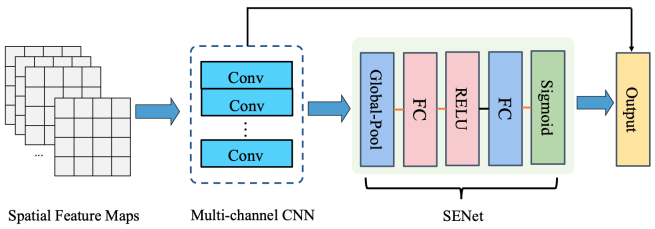

Fig. 4: The structure of Channel-wise CNN Block.

external features and the traffic accident risk spatial maps as the input and outputs the high-level representations. Specifically, in time slot t, the multi-channel features are represented as a tensor $M^t \in \mathbb{R}^{D\times H\times W}$, where D is the number of feature channels. In order to make SENet work better, we use a convolution layer to reconstruct the dimension of input channels. We compress the channel dimension from D to C. Then the results are fed into SENet to obtain the channel-wise features. The multi-channel CNN can be further formulated as the following equation:

$$M^{t,k} = \varphi(W^{t,k} * M^{t-1,k}) + b^{t,k} \tag{3}$$

where k is the layer of the convolution network, $W^{t,k}, b^{t,k}$are trainable weights; φ is the ReLU activation function, $*$ is the convolution operate, and $M^{t,0} = M^t$.

After l convolution layers, the output is denoted as $M^{t,l} = [m_1, m_2, ..., m_c] \in \mathbb{R}^{C\times H\times W}$. Each element m_i represents the learned feature embedding on each channel. For SENet, we first use the global average pooling to generate channel-wise statistics s as follows.

$$s = \frac{1}{H\times W} \times \sum_{i=1}^{H}\sum_{j=1}^{W} m(i,j) \tag{4}$$

With the channel-wise statistics as the channel importance coefficient, we next use the following equation to learn the nonlinear relationship between different channels by forming a bottleneck with two fully-connected (FC) layers.

$$z = \sigma(g(s,W)) = \sigma(W_2 * ReLU(W_1 * s)) \tag{5}$$

where W_1, W_2 are the parameters of fully-connected (FC) layers, $W_1 \in \mathbb{R}^{\frac{C}{r}\times C}, W_2 \in \mathbb{R}^{C\times\frac{C}{r}}$, and σ is the Sigmoid activation function. We next multiply the learned weight coefficient of each channel with the original representation to obtain a new representation, and add it with the initial input as the final output: $U_t \in \mathbb{R}^{C\times H\times W}$.

$$U^t = M^{t,l} + z * M^{t,l} \tag{6}$$

3.3 Multi-View Self-Attention GCN for Global Semantic Feature Learning

We next introduce the proposed multi-view self-attention graph convolution network module to capture the global semantic dependencies from the context features. We will first present how to construct the multi-view spatial graph, and then introduce the multi-view self-attention GCN.

3.3.1 Multi-View Semantic Graph Construction

The spatial correlations of the traffic accident risks in different regions are complex including both local and global spatial correlations. Two cell regions with similar POI distributions or functions (e.g. commercial areas) but geographically far away may also present similar traffic accident occurrence patterns. The global semantic correlation of the traffic accident risks can be largely reflected by the land features such as POIs and road networks [17].

To capture the global spatial correlations reflected by land features, we construct a multi-view spatial graph by modeling the global semantic relevance among cell regions from the perspectives of POI similarity, road feature similarity and traffic accident risk similarity. Note that each node in the graph is corresponding to a pixel in the spatial map. The POI distribution of a cell region can reflect the function of the region as well as the traffic flow patterns. Road features including the length of roads, road type, number of lanes in road and the numbers of overhead electronic signs can also reflect the traffic pattern of a cell region.

To measure the POI and road feature similarity between two cell regions, we calculate their distribution similarity based on Jensen-Shannon (JS) divergence [21]. JS divergence is widely used to measure the similarity of two probability distributions. For example, we compute the similarity of the POI distributions $P_{i,j}$ and $P_{m,n}$ between two cell regions $r_{i,j}$ and $r_{m,n}$.The POI distribution similarity between the two cell regions can be calculated as follow:

$$\Phi(P_{i,j}, P_{m,n}) = 1 - JS(P_{i,j}, P_{m,n}) \tag{7}$$

$$JS(P_{i,j}, P_{m,n}) = \tfrac{1}{2}\sum_{k=1}^{K}(P_{i,j}(k)\log\frac{2P_{i,j}(k)}{P_{i,j}(k)+P_{m,n}(k)+\epsilon} + P_{m,n}(k)\log\frac{2P_{m,n}(k)}{P_{i,j}(k)+P_{m,n}(k)+\epsilon}) \tag{8}$$

where K is the number of POI types. Based on the POI similarity and road feature similarity among all the cell regions, we can construct the two-view spatial graph $\mathcal{G}(\mathcal{V},\mathcal{E})$ that contains the POI similarity graph $G^p(\mathcal{V},E^p)$ and the road feature similarity graph $G^f(\mathcal{V},E^f)$. For each pair of cell regions $r_{i,j}$ and $r_{m,n}$, we also compute their traffic accident risk similarity with the DTW algorithm. The historical traffic accident risks of the two regions form two time series, and DTW algorithms can be used to measure the similarity of the two time series data. For each cell region $r_{i,j}$, we find the Top-k cell regions with the highest traffic accident risk similarity, and construct an edge between $r_{i,j}$ and the k most similar nodes. In this way, we construct the traffic accident risk graph $G^r(\mathcal{V},E^r)$.

By incorporating the traffic accident risk graph $G^r(\mathcal{V},E^r)$, the final multi-view spatial graph $\mathcal{G}(\mathcal{V},\mathcal{E})$ contains the POI similarity graph $G^p(\mathcal{V},E^p)$, the road feature similarity graph $G^f(\mathcal{V},E^f)$ and the traffic accident risk graph $G^r(\mathcal{V},E^r)$.

3.3.2 Multi-View Self-Attention GCN

To learn the global semantic features of the nodes in the constructed multi-view spatial graph, a multi-view self-attention GCN is proposed. In order to perform GCN in the spectral domain, we first calculate Laplacian matrix L as follows.

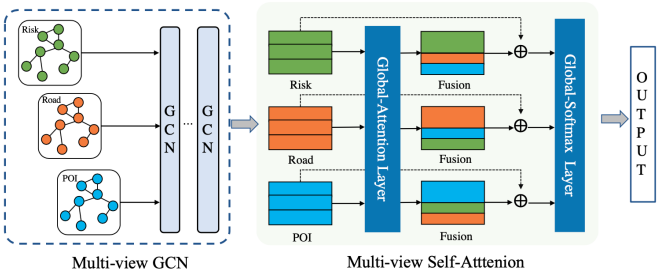

Fig. 5: The structure of multi-view self-attention GCN.

In time slot t, the graph signal matrix can be denoted as $G^t \in \mathbb{R}^{d_g \times N}$, where d_g is node features, N is the number of nodes, and A is the adjacent matrix. First, we derive O as follows

$$O = A + I_m \tag{9}$$

where I_m is the identity matrix of $N \times N$. Next, we can obtain the diagonal matrix D based on O

$$D = \begin{bmatrix} D_{11} & 0 & ... & 0 \\ 0 & D_{22} & ... & 0 \\ . & ... & ... & . \\ . & ... & ... & . \\ 0 & ... & ... & D_{NN} \end{bmatrix} \tag{10}$$

where $D_{ii} = \sum_{j=1}^{N} O_{ij}$, O_{ij} is the element in matrix O. Then we can get the Laplacian matrix L

$$L = (D^{-\frac{1}{2}})O(D^{-\frac{1}{2}}) \tag{11}$$

Next, we employ a two-layer GCN for node representation learning on each view graph of $\mathcal{G}$ as shown in the left part of Fig. 5. The graph convolution operation on the two layers can be formulated as follows.

$$\mathcal{H}^{t,n+1} = ReLU(L\mathcal{H}^{t,n}W^n + b^n), \tag{12}$$

$$\mathcal{H}^{t,0} = G^t, \tag{13}$$

where W^n and b^n are trainable parameters, $L \in \mathbb{R}^{N \times N}$ is the Laplacian matrix, and G^t is the graph signal matrix. With the above operations over the multi-view spatial graph at time slot t, we obtain the graph embedding $\mathcal{H}^t(G^r), \mathcal{H}^t(G^f), \mathcal{H}^t(G^p) \in \mathbb{R}^{N \times d_h}$ for the three views, respectively. For simplicity, we denote the graph embedding of the three views as $\mathcal{H}^t = [\mathcal{H}^1, \mathcal{H}^2, \mathcal{H}^3]$. Through the global-attention layer, each view can have a new representation that incorporates cross-view information as shown in the middle of Fig. 5. In order to preserve the original information, the original embedding and the new embedding are integrated through a fusion gate. Finally, we use the fusion result as the input of the global softmax layer to capture the different effects of three views on the risk prediction as shown in the right of Fig. 5.

Specifically, given input $\mathcal{H}^t = [\mathcal{H}^1, ..., \mathcal{H}^n] \in \mathbb{R}^{n \times N \times d_h}$, where the n is the number of views, N is the number of nodes, d_h is the graph convolution embedding size. Inspired by self-attention mechanism [22], three subspaces are obtained with $\mathcal{H}$, namely, query matrix Q, value matrix $V = \mathcal{H}$ and a key matrix K as follows:

$$K = \mathcal{H}W_k, \tag{14}$$

$$Q = \mathcal{H}W_q \tag{15}$$

Then, we spread the information among all the three views as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{16}$$

In addition, in order to get richer global information for each view, we propose the multi-view attention based on multi-head attention for each view. Specifically, we concatenate the features of three views together and project again to get the final values. The process can be formulated as the following equations:

$$MultiView(Q, K, V) = Concat(view_1, ..., view_n)W_o, \tag{17}$$

$$view_i = Attention(Q_i, K_i, V_i) \tag{18}$$

where $Q_i = \mathcal{H}W_i^Q, K_i = \mathcal{H}W_i^K, V_i = \mathcal{H}W_i^V$, and W_o is the trainable parameter. After the multi-head attention layer, we can get the relevant global information for each view: $\mathcal{H}^{'} = [\mathcal{H}^{1'}, \mathcal{H}^{2'}, ..., \mathcal{H}^{n'}] = MultiView(Q, K, V))$. Then we employ a residual connection in order to preserve the original information by gated linear units, which can be calculated for i-th view as follows:

$$\mathcal{H}^{i,new} = \alpha\mathcal{H}^i + (1 - \alpha)\mathcal{H}^{i'} \tag{19}$$

The final output can be obtained with a global-softmax layer as follows.

$$\widetilde{\mathcal{H}} = softmax(\mathcal{H}^{1,new}, ..., \mathcal{H}^{n,new}) \tag{20}$$

3.4 Cross-Scale GCN for Fine- and Coarse-Grained Feature Fusion

Different from video data, there are complex spatio-temporal correlations between regions in both fine- and coarse-grained traffic networks. For example, the nodes in the coarse-grained traffic network may cover commercial areas, schools, or other important POIs. The semantic information of the POIs in the coarse-grained traffic network can be very useful in transportation planning, and thus should be captured during both coarse- and fine-grained feature learning.

As shown in Fig. 2, the multi-view GCN is conducted on coarse- and fine-grained traffic accident risk data separately. To integrate the feature representations of the two granularities, we need to correlate the multi-view GCN between fine- and coarse-grained prediction tasks. As each coarse-grained cell region is composed of several fine-grained cell regions, the embedding of a coarse-grained region should be regarded as the aggregation of the embedding of the corresponding fine-grained regions. In short, for a coarse-grained cell region, each fine-grained region in it should reflect its partial spatial features. Therefore, the aggregation of the fine-grained cell region features should be consistent with the corresponding coarse-grained cell region features. To this end, we use an assignment matrix $B_{fc} \in \mathbb{R}^{N \times \frac{N}{k}}$

as follows to store the correspondence between fine- and coarse-grained cell regions.

$$B_{fc}(i,j) = \begin{cases} 1 & \text{if } r_i^f \in r_j^c \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

Then, in time interval t, assume that coarse- and fine-grained graph embeddings are $\widetilde{\mathcal{H}}_c^t \in \mathbb{R}^{\frac{N}{K} \times d_h}$ and $\widetilde{\mathcal{H}}_f^t \in R^{N \times d_h}$. We propose a cross-scale GCN module as follows to share the message between the two scales data as follows.

$$\overline{\mathcal{H}_c^t} = \widetilde{\mathcal{H}_c^t} + \sigma(B_{fc}^T \widetilde{\mathcal{H}_f^t} W_c) \tag{22}$$

$$\overline{\mathcal{H}_f^t} = \widetilde{\mathcal{H}_f^t} + \sigma(B_{fc} \widetilde{\mathcal{H}_c^t} W_f) \tag{23}$$

where W_f, W_c are the trainable parameters, B_{fc} is the assignment matrix and B_{fc}^T is its transposed matrix. With the above cross-scale GCN operations, the feature representation learning on fine- and coarse-grained graphs can be effectively fused and integrated.

3.5 Attention LSTM For Temporal Feature Learning

With the spatial and semantic features learned by the above methods, next we introduce how to learn the temporal features. Here we propose to use attention LSTM model to learn the temporal correlations of the historical traffic accident risks. Specifically, taking the fine-grained result $\overline{\mathcal{H}_f^t} \in \mathbb{R}^{N \times d_h}$ in time slot t after multi-view graph convolution as an example, the LSTM module as follows is first used to learn the temporal correlations of the features in successive time slots.

$$f_t = \sigma(W_f[h_{t-1}, \overline{\mathcal{H}_f^t}] + b_f) \tag{24}$$

$$i_t = \sigma(W_i[h_{t-1}, \overline{\mathcal{H}_f^t}] + b_i) \tag{25}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_c[h_{t-1}, \overline{\mathcal{H}_f^t}] + b_c) \tag{26}$$

$$o_t = \sigma(W_o[h_{t-1}, \overline{\mathcal{H}_f^t}] + b_o) \tag{27}$$

$$h_t = o_t \odot tanh(c_t) \tag{28}$$

where f_t is the output of forget gate, i_t and o_t are the output of input gate and output gate, c_t and h_t are the cell output and hidden state of a cell. Then for the output of hidden state of a cell h_t, attention mechanism is applied as follows.

$$Q = W_q h_t \tag{29}$$

$$K_t = W_k h_t \tag{30}$$

$$\alpha_t = K_t^T Q \tag{31}$$

$$Attention = Softmax(\alpha_1, ..., \alpha_m) \tag{32}$$

$$\mathcal{P}_{graph}^f = \sum_{t=1}^{m} h_t Attention(t) \tag{33}$$

where m is the number of time step, W_q and W_k are the trainable parameter. $\mathcal{P}_{graph}^f$ is the feature representation of the multi-view graph which fuses the global semantic and temporal features.

3.6 Local Spatial and Global Semantic Features Fusion

Through the Attention LSTM, taking the fine-grained features for example, it outputs two types of feature representations. The first is the local spatial and temporal features representation, which can be denoted as $\mathcal{P}_{grid}^f$. The second is the global semantic and temporal features representation, which can be denoted as $\mathcal{P}_{graph}^f$. Next, we need to integrate the two types of features to fuse the local spatial and global semantic features. We use a fully-connected layer as follows to fuse the two types of features.

$$X^{f^{T+1}} = FC(W_1 \mathcal{P}_{graph}^f + W_2 \mathcal{P}_{grid}^f). \tag{34}$$

where $FC(\cdot)$ denotes fully-connected layer, and W_1,W_2 are the learnable parameters.

4 Objective function for Multi-task Learning

Our goal is to predict the fine- and coarse-grained traffic accident risks simultaneously. Thus the two prediction tasks can be conducted jointly under a multi-task learning framework. The objective function of the designed multi-task learning model contains three parts: the prediction loss for the fine-grained data, the prediction loss for the coarse-grained data and a structure constraint loss between two views.

$$Loss = Loss_f + \lambda_1 Loss_c + \lambda_2 L_{fc} \tag{35}$$

where λ_1, λ_2 are the parameters of the loss function which can balance the importance of the three losses. $Loss_f$ and $Loss_c$ can be formulated as follows.

$$Loss_f = \frac{1}{N_f} ||X^f - \hat{X}^f||^2 \tag{36}$$

$$Loss_c = \frac{1}{N_c} ||X^c - \hat{X}^c||^2 \tag{37}$$

where $X^f \in \mathbb{R}^{N_f \times 1}$ and $X^c \in \mathbb{R}^{N_c \times 1}$ are the ground truth of the fine- and coarse-grained traffic accident risks, $\hat{X}^f \in \mathbb{R}^{N_f \times 1}$ and $\hat{X}^c \in \mathbb{R}^{N_c \times 1}$ are the predictions, and N_c, N_f are the numbers of the coarse- and fine-grained cell regions, respectively.

L_{fc} is a structure constraint loss between the two scales of data. Based on our definition on cell regions, a coarse-grained cell region contains $k \times k$ fine-grained cell regions, where k is the coarsening coefficient. To represent the relationship between the fine- and coarse-grained cell regions, we use an assignment matrix B_{fc}, whose each entry $B_{fc}(i,j)$ denotes the i-th fine-grained region belonging to the j-th coarse-grained region (value 1) or not (value 0) based on formula (21). With such a correspondence relationship between fine- and coarse-grained cell regions, the traffic accident risks of a coarse-grained regions should be consistent with that of the corresponding fine-grained regions. Namely, the traffic accident risks should be consistent with the summation of the traffic accident risks of all the fine-grained cell regions that belongs to the coarse-grained region. By taking this structure constraint into consideration, we design a structure constraint as follows.

$$L_{fc} = \frac{1}{N_c} ||\hat{X}^c - \hat{X}^f B_{fc}||^2 \tag{38}$$

where $\hat{X}^f B_{fc}$ is the inferred coarse-grained prediction based on the aggregation of the fine-grained predictions. The goal of this constraint term is to make the coarse-grained prediction close to the aggregation of the corresponding fine-grained predictions.

5 EXPERIMENT

In this section, we will conduct extensive experiments to evaluate the performance of MVMT-STN. We will first introduce the experiment settings including the datasets, the evaluation metrics, model implementation details and baselines. Then we will present and discuss the experiment results. Ablation study and hyperparameter study will be also conducted to test whether all the proposed components of our model are useful and how sensitive our proposal is to the hyperparameters. Finally, a prediction result visualization is given.

5.1 Datasets

As shown in Table 1, we conduct experiments on two large traffic accident datasets collected from two cities, New York City (NYC for short) and Chicago. The time span of NYC data is one year from January 2013 to December 2013. The time span of Chicago data is 8 months, from February 2016 to September 2016. For each dataset, besides the traffic accident data, we also use the taxi trip data as human mobility data, land features and external features. For NYC data, we also use Points of Interest (POI) data, road network features, taxi trips and weathers. For Chicago data, we have taxi trips, weathers and road features. POI data contains seven categories: residential areas, school, cultural facilities, entertainment, social services, transportation and commerce. We use the taxi trips to calculate the inflow and outflow in each cell region as the human mobility data. The weather data includes temperature and specific conditions including sunny, rainy, cloudy, snowy and mist. Road features contain road types, road segment length and width, and snow removal priority. The details of the two datasets are shown in Table 1.

TABLE 1: Statistics and description of the two datasets.

City	Dataset	Time Span	# of Records
NYC	Accidents	01/01/2013 - 12/31/2013	147k
	Taxi Trips		173,179k
	POIs		15,625
	Weathers		8,760
	Road Network		103k
Chicago	Accidents	2/1/2016 - 9/30/2016	44k
	Taxi Trips		1,744k
	Weathers		5,832
	Road Network		56k

5.2 Evaluation Metrics

We use three metrics to evaluate the performance of our model. By considering our prediction task as a regression problem, Root Mean Square Error (RMSE) is used. Inspired by [12, 19], the studied task can be also considered as a classification problem, and thus we can use Accuracy of top L (Acc @ L) and mean average precision of L(MAP) to measure the prediction accuracy. $Acc@L$ is the percentage of accurate predictions for a prediction list of length L, and MAP is widely used as the global evaluation of ranking tasks. These metrics are defined as follows:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}||X_t - \hat{X}_t||^2} \tag{39}$$

$$Acc = \frac{1}{T}\sum_{t=1}^{T}\frac{P_t \cap R_t}{|R_t|} \tag{40}$$

$$MAP = \frac{1}{T}\sum_{t=1}^{T}\frac{\sum_{j=1}^{|R_t|} pre(j) \times real(j)}{|R_t|} \tag{41}$$

where $\hat{X}^t$ is the prediction in time slot t, X^t is the ground truth, P_t is the ground truth set of areas where the traffic accident risks are in the top L, and R_t is the set of areas where the predicted risks are in the top L. $pre(j)$ indicates the precision of a cut-off rank list from 1 to j. $real(j)$ is 1 when there are accidents occur in this area. In fine- and coarse-grained prediction, L is set to 50 and 20, respectively, which means that the top-50 and 20 areas with the highest risk in fine- and coarse-grained cell regions will be considered as high-risk areas.

5.3 Implementation Details

In our experiment, the data is divided into training, validation and test set at a ratio of 6:2:2. The entire city area is divided into 20×20 cell regions. Since not all regions have a road network structure, 243 cell regions with road network structures are preserved in the fine-grained prediction of New York City and 75 regions are preserved for a coarse-grained prediction. For Chicago, there are 197 cell regions with road network structures for fine-grained prediction and 59 coarse-grained regions. In order to make the training more effective, all data are normalized with the Max-Min normalization method before training. The length of long-term (weekly) ρ and short-term (recently) κ are set to 3 and 4, respectively. For GCN in each view, we stack two layers graph convolution operations and the number of kernels in each layer is 64. For the channel-wise CNN block module, the convolution kernel is set to 3×3. The learning rate and batch size are set to 10^{-4} and 32, respectively. Part of the model is implemented with Huawei MindSpore framework and the experiment is conducted on the GPU of NVIDIA GeForce RTX 3090. The code is publicly available at github https://github.com/ZJ-LEARN/MVMT-STN.

5.4 Baselines

We compare MVMT-STN with the following baselines.

•**ARIMA**: ARIMA is a typical time series forecasting model, which consists of three parts: AR model (autoregressive model), MA model (moving average model), and the order of difference.

• **LSTM** [23]: LSTM is a classic deep learning-based time series forecasting model, which can capture the nonlinear dependencies between data in the time dimension for traffic accident risk prediction.

• **MLP**: MLP is a neural network composed of fully connected layers, and the output of each hidden layer is transformed by an activation function, which has adaptive learning capabilities. The activation function is ReLU.

• **ConvLSTM** [24]: ConvLSTM is a method that effectively combines convolution and LSTM. Input to state and state-to-state conversion of ConvLSTM have a convolution structure, which can effectively process spatio-temporal data.

TABLE 2: Performance comparison among different methods

Model	NYC			Chicago		
	RMSE (F/C)	Acc@50 (F)/Acc@20 (C)	MAP (F/C)	RMSE (F/C)	Acc@50 (F)/Acc@20 (C)	MAP (F/C)
ARIMA	0.3952/0.7795	29.39%/36.3%	0.0431/0.1493	0.338/0.461	42.65%/54.23%	0.0329/0.1413
MLP	0.4007/0.7535	34.35%/46.59%	0.1264/0.2979	0.2346/0.4221	41.47%/57.13%	0.0476/0.2154
LSTM	0.3979/0.7404	34.9%/48.42%	0.1033/0.2910	0.2305/0.4282	42.99%/57.51%	0.0596/0.2203
ConvLSTM	0.393/0.7438	36.22%/53.86%	0.1326/0.3343	0.214/0.4148	44.68%/62.13%	0.065/0.2251
H-ConvLSTM	0.3912/0.7376	39.2%/54.10%	0.161/0.3428	0.2113/0.4169	45.71%/60.4%	0.0736/0.2405
STGCN	0.3923/0.7187	43.19%/54.35%	0.1753/0.3423	0.211/0.4137	49.25%/63.87%	0.0858/0.2677
T-GCN	0.3849/0.7261	44.07%/55.88%	0.1766/0.3647	0.2103/0.4008	50.31%/64.41%	0.0823/0.2663
GSNet	0.3958/0.7347	46.79%/57.67%	0.1832/0.3741	0.2103/0.4033	50.4%/65.83%	0.0922/0.2838
MVMT-STN	**0.3829/0.7183**	**48.59%/58.75%**	**0.1977/0.3835**	**0.2095/0.4023**	**53.71%/69.52%**	**0.1067/0.2852**

TABLE 3: Performance comparison among different methods in rush hours (7:00-9:00 am and 16:00-19:00 pm)

Model	NYC			Chicago		
	RMSE (F/C)	Acc@50 (F)/Acc@20 (C)	MAP (F/C)	RMSE (F/C)	Acc@50 (F)/Acc@20 (C)	MAP (F/C)
ARIMA	0.4645/0.8873	26.60%/36.53%	0.0747/0.317	0.3143/0.5255	39.20%/48.94%	0.0523/0.2203
MLP	0.4522/0.8688	34.33%/45.40%	0.1319/0.3504	0.2751/0.5269	39.84%/54.76%	0.0639/0.2858
LSTM	0.4443/0.8479	34.19%/47.53	0.1151/0.3480	0.2709/0.5086	40.56%/55.23%	0.0737/0.2701
ConvLSTM	0.4422/0.855	35.65%/51.05%	0.1266/0.369	0.2665/0.5207	42.34%/59.39%	0.0894/0.2848
H-ConvLSTM	0.4399/0.8371	37.05%/52.26%	0.153/0.3915	0.2632/0.5235	43.23%/57.17%	0.0830/0.3167
STGCN	0.443/0.8207	40.84%/52.56%	0.1694/0.3935	0.263/0.5192	46.19%/61.89%	0.1007/0.3256
T-GCN	0.4318/0.8195	42.71%/54.28%	0.1684/0.3985	0.2603/0.4966	47.83%63.46%	0.1069/0.3349
GSNet	0.4382/0.8248	44.62%/55.67%	0.1747/0.3955	0.2597/0.5013	48.38%/63.55%	0.1132/0.3363
MVMT-STN	**0.4335/0.8156**	**46.58%/56.60%**	**0.1902/0.4010**	**0.2593/0.5025**	**51.06%/66.53%**	**0.1330/0.3587**

• **Hetero-ConvLSTM** [7]: It is a deep learning model proposed for traffic accident prediction, which uses spatial features and spatial model integration to solve the challenge of spatial heterogeneity.

• **STGCN** [25]: It is a graph based deep learning model, which combines graph convolution and gated time 1-D convolution for spatio-temporal traffic flow prediction.

• **T-GCN** [26]: It is a traffic flow forecasting method based on neural network, which combines GCN and GRU to extract spatio-temporal features.

• **GSNet** [19]: It is a recently proposed framework which can effectively capture the complex spatio-temporal correlation from both geographic and semantic aspects for traffic accident risk prediction. A weighted loss function is designed to solve the zero inflation problem.

5.5 Experiment Result

5.5.1 Performance Comparison

Tables 2 shows the prediction performance under the three evaluation metrics of MVMT-STN and baselines over the two datasets. To further test the model performance in rush hours when traffic accidents are more likely to occur, we also show the prediction performance of these methods in rush hours from 7:00 to 9:00 am and from 16:00 to 19:00 pm in Table 3. It shows that MVMT-STN consistently achieves the best results in all the cases. One can also see that with the multi-task learning framework, both the fine- and coarse-grained prediction performance is improved, which verifies the effectiveness of the proposed multi-task learning model. Specifically, on Chicago dataset, MVMT-STN improves the performance achieved by the best baseline by 15.7% and 17.4% on MAP for the two time periods. For NYC dataset, the performance improvements are 7.9% and 8.8% on MAP for the two time periods. It reveals that MVMT-STN can better capture the causes of traffic accidents in different situations , so it can better predict the risk coefficient ranking of each region. One can see that RMSE value of coarse-grained prediction is always larger than that of fine-grained prediction. This is mainly because the coarse-grained risk prediction is the sum of the predicted risks of the corresponding fine-grained areas. Since the fine-grained area is much larger and the traffic accident risk is also higher than the fine-grained area, the prediction error for coarse-grained prediction is larger. The coarse-grained prediction results can also reflect the effectiveness of MVMT-STN in coarse-grained risk modeling and the scalability of MVMT-STN in city-wide traffic accident risk forecasting.

The performance of ARIMA is inferior to deep learning-based methods because it cannot capture the complex non-linear spatio-temporal correlations. MLP is a relatively simple neural network model and is less effective to mine sufficient information. LSTM only considers the temporal correlation, but ignores the spatial correlation. Thus the perfor-

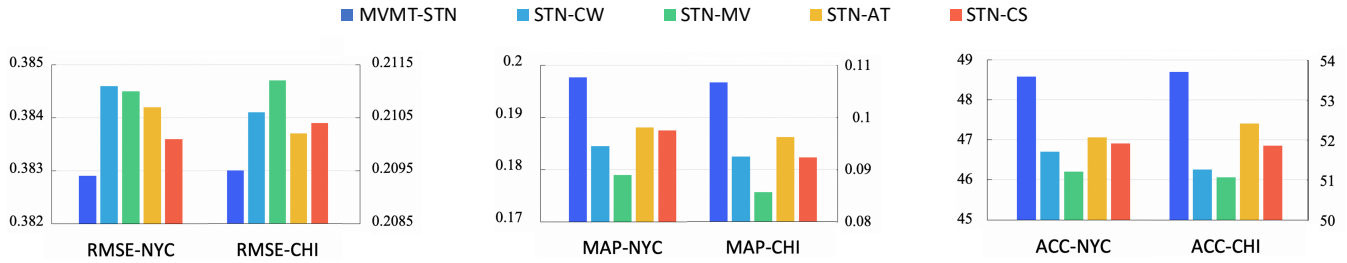

Fig. 6: Performance comparison between MVMT-STN and its variant models.

mance of these simple models are all inferior to other deep learning models. ConvLSTM considers the temporal and spatial correlation simultaneously while Hetero-Convlstm also takes the external factors into account. Thus one can see that H-ConvLSTM performs better than ConvLSTM, which verifies that the external factors are useful for the studied problem. Graph-based models including STGCN, T-GCN perform better than CNN-based models ConvLSTM and H-ConvLSTM, indicating that the road network features and the global semantic features are also helpful to improve the prediction performance. GSNet considers both the global semantic and local geographic spatial features. However, it ignores the diverse impacts of various external factors on traffic accidents. Overall, our model MVMT-STN can effectively capture the diverse impacts of different external factors on traffic accidents while taking the complex temporal and spatial correlations into account, and thus achieves the best performance among all the methods.

5.5.2 Ablation Study

In order to investigate whether all the components designed in our model are all useful to the studied problem, we conduct ablation study in this subsection. We remove Channel-wise CNN block, Multi-view GCN, attention, and structure constraint components respectively from MVMT-STN as four variant models. We name the four variants as STN-CW, STN-MV, STN-AT and STN-CS. Fig. 6 shows the RMSE, Acc@50, MAP comparison between MVMT-STN and the four variants in fine-grained prediction. One can see that MVMT-STN outperforms the four variants under all the three metrics, indicating the effectiveness of the proposed four components. One can also observe that the multi-view GCN contributes most to the model performance improvement with up to 7.2% improvement in terms of MAP. It indicates the multi-view global semantic correlation is especially important to help the task of traffic accident risk prediction. The prediction performance drops significantly when the structure constraint loss is removed, which demonstrates the multi-task learning between fine- and coarse-grained prediction does improve the performance compared with single-task learning. Removing the attention mechanism of LSTM also hurts the performance, which proves the necessity of dynamically modeling the importance of historical data. Combining these components together achieves the best performance. Thus one can conclude that the well-designed components in MVMT-STN are also useful for the traffic accident risk prediction problem.

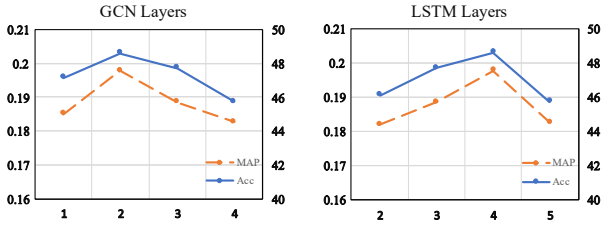

Fig. 7: Performance with different GCN and LSTM layers

5.5.3 Hyperparameter Study

In order to study the influence of different hyperparameters on the mode performance, we show the performance results under different number of GCN and LSTM layers in Fig. 7 on the NYC dataset. One can observe from the figure that the model achieves the best performance when the number of GCN layers is 2 and LSTM layer number is 4. Too small (1 layer) or too many (3 or 4 layers) layers will hurt the performance. This is probably because only 1 layer GCN cannot fully capture the complex spatial correlations, while too many layers will lead to overfitting. A proper number like 4 of LSTM layers can better capture the long-term temporal correlation. Next, we study the effect of λ_1 and λ_2 in the final objective function of multi-task learning to the model performance. We tune the values of λ_1 and λ_2 and record the RMSE, Acc@50 and MAP under different value settings. The result is shown in Table 4. One can see that the three metrics vary with different values of the two parameters, which means they both have significant impact on the model performance. The best performance is achieved when $\lambda_1 = 1.4$ and $\lambda_2 = 0.4$.

TABLE 4: Performance on different loss weight settings

λ_1	λ_2	RMSE	Acc@50	MAP
0.4	0.4	0.394	45.23	0.1658
1	1	0.3926	46.46	0.1842
1.4	0	0.3956	46.29	0.1821
1.4	0.4	0.3829	48.59	0.1977
1.4	1	0.3839	47.26	0.1932
1.4	1.4	0.3912	47.67	0.1867

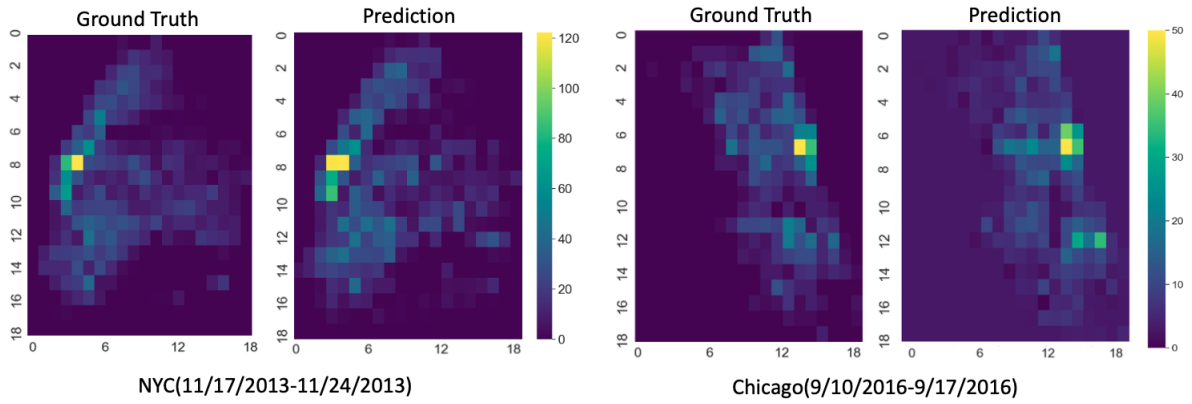

Fig. 8: Fine-grained visualization for traffic accident risk forecasting on two datasets

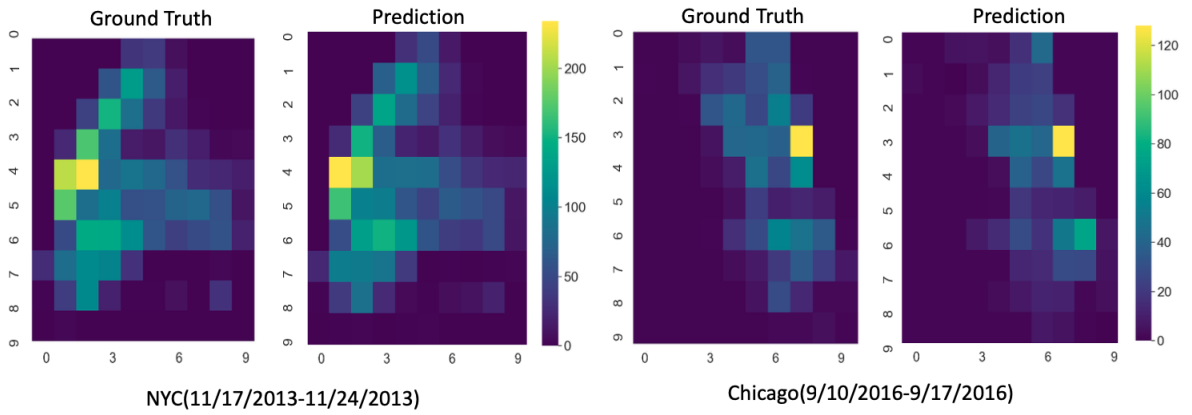

Fig. 9: Coarse-grained visualization for traffic accident risk forecasting on two datasets

5.5.4 Visualization

To further test the ability of our model in traffic accident risk prediction, we visualize the prediction result and the ground truth on the two datasets. Fig. 8 and Fig. 9 show the visualization result for the predicted fine- and coarse-grained traffic accident risk heat map and the ground truth in the two cities within a week from 11/17/2013 to 11/24/2013 for NYC and from 9/10/2016 to 9/17/2016 for Chicago. It can be observed that overall the predicted heat maps of traffic accident risks on both datasets are very similar to the ground truth, which means that our model can provide a very accurate prediction. One can also observe that the traffic accident data are truly sparse as most regions have a very small number of traffic accidents, while only a small proportion of regions (central areas of a city) have high traffic accident risks. For the central areas of the city where the traffic flow is high, our prediction is also very accurate as the predicted heat map color is very close to the ground truth color. Thus the visualization result further verifies the effectiveness of the proposed MVMT-STN model.

5.5.5 Efficiency Analysis of MVMT-STN

We show the computation time of MVMT-STN and baseline methods in Fig. 10 to study the computational efficiency of these methods. The left y-axis indicates the time and the right y-axis indicates the parameter size. One can see that benefiting from the temporal convolution structure, STGCN has the least number of parameters and thus it needs much less training time. As MVMT-STN needs to predict both coarse- and fine-grained traffic accident risks, it needs more training time. Thus MVMT-STN needs less training time than GSNet and Hetero-ConvLSTM, but needs more time than other methods. Although MVMT-STN needs more training time, the prediction is pretty fast, only needing around 2.58 seconds to complete the multi-scale traffic accident risk prediction for the entire city, which can meet the requirement of real-time forecasting.

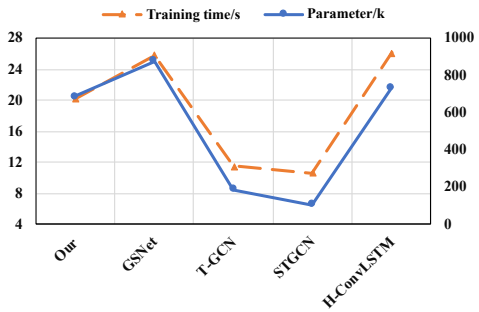

Fig. 10: The computation time and parameter size comparison

6 Related work

We review related work from the following two aspects that are highly relevant to this work: traffic accident prediction and spatio-temporal data prediction with deep learning.

Traffic accident prediction. The prediction of urban traffic accidents is one of the key topics of intelligent transportation systems. Existing solutions can be generally categorized into two types, conventional pattern-based methods and deep learning-based models. Conventional pattern-based methods include SVM, decision tree, KNN, etc. Bharti et al. [27] used support vector machines (SVM) with Gaussian kernel to analysis the urban traffic accident. Caliendo et al [28] used Poisson, negative binomial and negative polynomial regression models of tangent and curve respectively to simulate the frequency of accidents. Lv et al. [16] selected traffic accident characteristics based on Euclidean distance, and used KNN method to complete the real-time prediction of road traffic accidents. Lin et al. [29] proposed a Frequent Pattern tree (FP tree) based variable selection method, and then utilized KNN and Bayesian Network to predict traffic accidents by using selected features. Pattern-based methods mostly rely on classical machine learning methods and stationary assumption to infer traffic accident. However, traffic accidents data have non-linear and complex spatial-temporal correlations, and the occurrence of traffic accidents can be significantly influenced by external context features including weather, POI and road feature. Without considering these features, the prediction accuracy of these methods is far from promising.

Recently, deep learning-based methods are widely used to predict traffic accidents. Yuan et al. [7] proposed a Hetero-Convolutional Long Short-Term Memory (Hetero-ConvLSTM) model to capture spatial heterogeneity and temporal auto-correlation of the traffic accidents. Ren et al. [30] extended the RNN architecture and Fully Connected (FC) network to predict the future traffic accident risk. Chen et al. [18] collected a large amount of heterogeneous data including traffic accident data and GPS records to understand the impact of human mobility on the risk of traffic accidents. Zhou et al. [12] used the differential Time-varying Graph neural network to obtain the influence of the traffic flow and spatial dependency. Wang et al. [19] proposed a model to capture both geo-spatial correlation and semantic spatial correlation. But they overlooked the diverse effect of various external context features.

Spatio-temporal data prediction with deep learning. The recent advances of deep learning techniques have greatly improved the research field of spatio-temporal data prediction [31, 32]. RNN and its variants including LSTM and GRU [24] can effectively model the nonlinear correlation in time dimension. The graph convolutional network (GCN) [25, 33] and convolutional Neural Network (CNN) [5] are widely used to model the spatial correlation on spatial images and graphs, respectively.

Deep learning are widely used in many spatio-temporal prediction applications, such as crowd flows prediction [34, 35, 36], crowd management [37], traffic accident prediction [25], air quality inference [38, 39], crime rate forecasting [40, 41] and land location recommendation [42] . In particular, Zhang et al. [5] use the residual neural network framework to model the time proximity, period and trend attributes of complex traffic. Yi et al. [39] proposed a model (Deep-Air) consisting of a spatial transformation component and a deep distributed fusion network, which can predict the air quality of each monitoring station in the future. Huang et al. [41] proposed the DeepCrime framework that can represent all space, time and type information together in a hidden vector, and at the same time use a hierarchical loop network to capture crime dynamics, thereby predicting the occurrence of different types of crime in each area. Zhao et al. [42] introduced context-aware embedding learning methods when studied POI recommendation tasks.

7 Conclusion

In this paper, we proposed a novel Multi-View Multi-Task Spatio-Temporal Network model named MVMT-STN to more effectively predict fine- and coarse-grained city-wide traffic accident risks simultaneously. Particularly, we adapted a multi-task learning framework to jointly forecast fine- and coarse-grained traffic accident risks to alleviate the data sparsity issue. A cross-scale GCN and a structure constraint loss were also designed to effectively couple the two prediction tasks over two data granularities. For each granularity, we proposed a channel-wise CNN and a multi-view GCN to capture both the local geographic and global semantic dependencies. In order to obtain the diverse impacts of the external context features on traffic accidents, we also proposed a feature fusion module which enabled weighted multi-channel fusion and cross-view information sharing. We evaluated our proposal over two real-world datasets, and the results verified that our model outperformed state-of-the-art methods.

In the future, it would be interesting to further study the spatio-temporal correlations of the traffic accident data in more than two scales, as finer grained correlation may further improve the model performance but also lead to sparser data in coarse-grained regions. It would be also interesting to study the extension of this model to other spatio-temporal data prediction tasks, such as urban crime prediction.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No.: 62172443), TRS-RGC Theme-based Research Scheme 2020/21 (No.: T41-603/20-R), CRF-RGC Collaborative Research Fund 2018/19 - Group Research Grant (RGC No.: C5026-18G), Guangdong Key Area R & D Plan (No. 2019B111106001), and CAAI-Huawei MindSpore Open Fund.

References

[1] Chinmoy Pal, Shigeru Hirayama, Sangolla Narahari, Manoharan Jeyabharath, Gopinath Prakash, and Vimalathithan Kulothungan. An insight of world health organization (who) accident database by cluster analysis with self-organizing map (som). *Traffic injury prevention*, 19(sup1):S15–S20, 2018.

[2] JD Tamerius, X Zhou, R Mantilla, and T Greenfield-Huitt. Precipitation effects on motor vehicle crashes

vary by space, time, and environmental conditions. *Weather, Climate, and Society*, 8(4):399–407, 2016.

[3] Lida Barba, Nibaldo Rodríguez, and Cecilia Montt. Smoothing strategies combined with arima and neural networks to improve the forecasting of traffic accidents. *The Scientific World Journal*, 2014, 2014.

[4] Honglei Ren, You Song, Jingwen Wang, Yucheng Hu, and Jinzhi Lei. A deep learning approach to the citywide traffic accident risk prediction. In *Proceedings of 21st International Conference on Intelligent Transportation Systems*, 2018.

[5] Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

[6] Chao Chen, Xiaoliang Fan, Chuanpan Zheng, Lujing Xiao, Ming Cheng, and Cheng Wang. Sdcae: Stack denoising convolutional autoencoder model for accident risk prediction via traffic big data. In *Sixth International Conference on Advanced Cloud and Big Data (CBD)*, 2018.

[7] Zhuoning Yuan, Xun Zhou, and Tianbao Yang. Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In *Proceedings of the 24th ACM SIGKDD*, 2018.

[8] Chao Huang, Chuxu Zhang, Peng Dai, and Liefeng Bo. Deep dynamic fusion network for traffic accident forecasting. In *Proceedings of 28th ACM International Conference on Information and Knowledge Management*, 2019.

[9] Jie Bao, Pan Liu, and Satish V Ukkusuri. A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data. *Accident Analysis & Prevention*, 122:239–254, 2019.

[10] Alameen Najjar, Shun'ichi Kaneko, and Yoshikazu Miyanaga. Combining satellite imagery and open data to map road safety. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

[11] Le Yu, Bowen Du, Xiao Hu, Leilei Sun, Liangzhe Han, and Weifeng Lv. Deep spatio-temporal graph convolutional network for traffic accident prediction. *Neurocomputing*, 423:135–147, 2021.

[12] Zhengyang Zhou, Yang Wang, Xike Xie, Lianliang Chen, and Hengchang Liu. Riskoracle: A minute-level citywide traffic accident forecasting framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[13] Ziqian Lin, Jie Feng, Ziyang Lu, Yong Li, and Depeng Jin. Deepstn+: Context-aware spatial-temporal neural network for crowd flow prediction in metropolis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[14] Ruth Bergel-Hayat, Mohammed Debbarh, Constantinos Antoniou, and George Yannis. Explaining the road accident risk: weather effects. *Accident Analysis & Prevention*, 60:456–465, 2013.

[15] Nicholas Jing Yuan, Yu Zheng, Xing Xie, Yingzi Wang, Kai Zheng, and Hui Xiong. Discovering urban functional zones using latent activity trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 27(3): 712–725, 2014.

[16] Yisheng Lv, Shuming Tang, and Hongxia Zhao. Real-time highway traffic accident prediction based on the k-nearest neighbor method. In *2009 international conference on measuring technology and mechatronics automation*, 2009.

[17] Yuxuan Liang, Zhongyuan Jiang, and Yu Zheng. Inferring traffic cascading patterns. In *Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems*, 2017.

[18] Quanjun Chen, Xuan Song, Harutoshi Yamada, and Ryosuke Shibasaki. Learning deep representation from big and heterogeneous data for traffic accident inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016.

[19] Beibei Wang, Youfang Lin, Shengnan Guo, and Huaiyu Wan. Gsnet: Learning spatial-temporal correlations from geographical and semantic aspects for traffic accident risk forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[20] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[21] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37 (1):145–151, 1991.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[23] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*, 2014.

[24] Xingjian Shi, Zhourong Chen, Hao Wang, Dit Yan Yeung, Wai Kin Wong, and Wang Chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 2015.

[25] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018.

[26] Ling Zhao, Yujiao Song, Chao Zhang, Yu Liu, Pu Wang, Tao Lin, Min Deng, and Haifeng Li. T-gcn: A temporal graph convolutional network for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems*, 21(9):3848–3858, 2019.

[27] Bharti Sharma, Vinod Kumar Katiyar, and Kranti Kumar. Traffic accident prediction model using support vector machines with gaussian kernel. In *Proceedings of fifth international conference on soft computing for problem solving*, 2016.

[28] Ciro Caliendo, Maurizio Guida, and Alessandra Parisi. A crash-prediction model for multilane roads. *Accident Analysis & Prevention*, 39(4):657–670, 2007.

[29] Lei Lin, Qian Wang, and Adel W Sadek. A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. *Transportation Research Part C: Emerging Technologies*, 55:444–459, 2015.

[30] Honglei Ren, You Song, JingXin Liu, Yucheng Hu, and Jinzhi Lei. A deep learning approach to the prediction

of short-term traffic accident risk.

[31] Senzhang Wang, Jiannong Cao, and Philip Yu. Deep learning for spatio-temporal data mining: A survey. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–20, 2020. doi: 10.1109/TKDE.2020.3025580.

[32] Senzhang Wang, Meiyue Zhang, Hao Miao, and Philip S. Yu. Mt-stnets: Multi-task spatial-temporal networks for multi-scale traffic prediction. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 504–512.

[33] Hao Peng, Hongfei Wang, Bowen Du, Md Zakirul Alam Bhuiyan, Hongyuan Ma, Jianwei Liu, Lihong Wang, Zeyu Yang, Linfeng Du, Senzhang Wang, and Philip S. Yu. Spatial temporal incidence dynamic graph neural networks for traffic flow forecasting. *Information Sciences*, 521:277–290, 2020.

[34] Senzhang Wang, Hao Miao, Hao Chen, and Zhiqiu Huang. Multi-task adversarial spatial-temporal networks for crowd flow prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020.

[35] Senzhang Wang, Jiannong Cao, Hao Chen, Hao Peng, and Zhiqiu Huang. Seqst-gan: Seq2seq generative adversarial nets for multi-step urban crowd flow prediction. *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, 6(4):1–24, 2020.

[36] Renhe Jiang, Zekun Cai, Zhaonan Wang, Chuang Yang, Zipei Fan, Quanjun Chen, Kota Tsubouchi, Xuan Song, and Ryosuke Shibasaki. Deepcrowd: A deep model for large-scale citywide crowd density and flow prediction. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[37] Renhe Jiang, Xuan Song, Dou Huang, Xiaoya Song, Tianqi Xia, Zekun Cai, Zhaonan Wang, Kyoung-Sook Kim, and Ryosuke Shibasaki. Deepurbanevent: A system for predicting citywide crowd dynamics at big events. In *Proceedings of the 25th ACM SIGKDD*, 2019.

[38] Weiyu Cheng, Yanyan Shen, Yanmin Zhu, and Linpeng Huang. A neural attention model for urban air quality inference: Learning the weights of monitoring stations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[39] Xiuwen Yi, Junbo Zhang, Zhaoyuan Wang, Tianrui Li, and Yu Zheng. Deep distributed fusion network for air quality prediction. In *Proceedings of the 24th ACM SIGKDD*, 2018.

[40] Bao Wang, Penghang Yin, Andrea Louise Bertozzi, P Jeffrey Brantingham, Stanley Joel Osher, and Jack Xin. Deep learning for real-time crime forecasting and its ternarization. *Chinese Annals of Mathematics, Series B*, 40 (6):949–966, 2019.

[41] Chao Huang, Junbo Zhang, Yu Zheng, and Nitesh V Chawla. Deepcrime: Attentive hierarchical recurrent networks for crime prediction. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018.

[42] Shenglin Zhao, Tong Zhao, Irwin King, and Michael R Lyu. Geo-teaser: Geo-temporal sequential embedding rank for point-of-interest recommendation. In *Proceedings of the 26th international conference on world wide web companion*, 2017.

Senzhang Wang received the B.Sc. and Ph.D. degree in Southeast University, Nanjing, China in 2009 and Beihang University, Beijing, China in 2016 respectively. He is currently a professor at School of Computer Science and Engineering, Central South University. His main research focus is on spatio-temporal data mining, graph data mining and urban computing. He has published more than 100 referred conference and journal papers including KDD, AAAI, IJCAI, CIKM, ICDM, SDM, TKDE, T-ITS, KAIS etc.

Jiaqiang Zhang received the B.s. degree in Information and Computing Science from Nanjing XiaoZhuang University, Nanjing, China, in 2020. He is currently a Master student of Nanjing university of aeronautics and astronautics in the department of Computer science and technology. From July 2019 to August 2019, he was a Visiting Student in University of California, Los Angeles, U.S.A. His research interest includes spatio-temporal data mining and deep learning.

Jiyue Li received the B.s. degree in Software Engineering from Guangzhou University, Guangzhou, China, in 2020. She is currently a Master student of Nanjing university of aeronautics and astronautics in the department of Computer science and technology. During July 2019, she was a Visiting Student in University of Washington, Seattle, U.S.A. Her research interest includes spatio-temporal data mining and deep learning.

Hao Miao received the B.s. degree in Computer science and technology from Nanjing Tech university, Nanjing, China, in 2018. He is currently a PhD student of Aalborg University in the department of computer science. From September 2019 to November 2019, he was a Visiting Student in The Hong Kong Polytechnic University, Hong Kong, China. His research interest includes spatio-temporal data mining, deep learning and transfer learning.

Jiannong Cao received the B.Sc. degree in computer science from Nanjing University, China, in 1982, and the M.Sc. and Ph.D. degrees in computer science from Washington State University, USA, in 1986 and 1990 respectively. He is currently a Chair Professor of Department of Computing at The Hong Kong Polytechnic University, Hong Kong. His research interests include parallel and distributed computing, wireless networks and mobile computing, big data and cloud computing, pervasive computing, and fault tolerant computing. He has co-authored 5 books in Mobile Computing and Wireless Sensor Networks, co-edited 9 books, and published over 500 papers in major international journals and conference proceedings. He is a fellow of IEEE, a distinguished member of ACM, a senior member of China Computer Federation (CCF).