Foresee Urban Sparse Traffic Accidents: A Spatiotemporal Multi-Granularity Perspective

Zhengyang Zhou, *Student Member, IEEE,* Yang Wang*, *Member, IEEE,* Xike Xie, *Member, IEEE,* Lianliang Chen, and Chaochao Zhu

Abstract—Traffic accident has become a significant health and development threat with rapid urbanizations. An accurate urban accident forecasting enables higher-quality police force pre-allocation and safe route planning for both traffic administrations and travelers, maximumly reducing injuries and damages. Off-the-shelf short-term accident forecasting methods, which focus on modeling static region-wise correlations with existing neural networks, mostly performed on hour levels and with single step. However, given the dynamic nature of road networks and expanding urban areas, it is challenging when the spatiotemporal granularity of forecasting improves as the rareness of accident records and complexity of long-term future dependencies. To address these challenges, we propose a unified framework RiskSeq, to foresee sparse urban accidents with finer granularities and multiple steps in spatiotemporal perspective. In particular, we design region-wise proximity measurements and temporal feature differential operations, and embed them into a novel Differential Time-varying Graph Convolution Network to dynamically capture traffic variations. Considering the hierarchical spatial dependencies and obvious context influences, a hierarchical sequence learning structure is devised by introducing contextual factors into a step-wise decoder. The multi-scale spatial risks are learned jointly to boost the risk predictions based on risk-gather and risk-assign networks. Extensive experiments demonstrate our RiskSeq can increase 5% to 15% performances on two datasets.

Index Terms—Traffic accident forecasting, spatiotemporal data mining, graph convolutional network, urban computing.

1 Introduction

Traffic accident has become into one of the biggest public health threats as World Health Organization (WHO) reported approximately 1.25 million people have died on roads during 2015 [1]. With constantly increasing number of vehicles, traffic accident forecasting is of great significance to reduce traffic injuries and ensure urban safety. For example, with some newly proposed models for predicting daily statewide accident risks, the fatality rate of traffic accidents in Tennessee has been reduced by 8.16% in 2016 [2]. Therefore, a spatiotemporal finer-grained and multi-granularity accident forecasting can not only benefit the public safety managements but also enhance the service qualities of various intelligent transportation systems, including real-time safe route recommendations for individual drivers and other location-based services.

There have been a wide range of researches delving into time-series predictions, inlcuding particle swarm optimization (PSO)-based [3], [4], [5] and ARIMA-based [6] methods. And general spatiotemporal predictions [7], [8], [9], [10], [11] have also been further studied. Nevertheless, all these existing works focus on continuous element forecasting. Regarding the issue of traffic accident forecasting, differing from those above-mentioned intensive and continuous predictions, it can be seen as a sporadic event forecasting. Specifically, traffic accident forecasting can be further classified into different categories with regard to the temporal granularities of long-term (daily predictions) and short-term (hourly or temporal finer-grained predictions) as well as the number of prediction steps, as summarized in Table 1. In particular, regarding long-term forecasting, methods like deep dynamic fusion network (DFN) [12], Hetero-ConvLSTM [13] and classfication-and-regression tree [13] were proposed to predict future daily risks by modeling the spatiotemporal heterogeneous data. However, these long-term forecasting approaches could not be directly used to address the more practical issue of real-time accidents predictions. To this end, early approaches for short-term accident predictions were proposed based on traditional machine learning [14], [15]. Nevertheless, none of these approaches have considered both the spatial and temporal correlations jointly. Recently, deep learning techniques including LSTM [16], autoencoder-based [17], and spatiotemporal attention-based [18], [19] were employed to address the challenging task by modeling citywide traffic risks during different periods as sequences, and these methods are all single temporal granularity and suffer the zero-inflated issue due to sporadic distributions of short-term accidents [26].

Unfortunately, predictions with single temporal step including both long-term traffic risk predictions and single-step short-term works, cannot independently support urban transportation applications since the durations of urban trips may be 15 minutes to hours in modern metropolis [20]. Further, traffic administrative agencies at different levels should have various spatial granularity requirements on predicting traffic risks due to their different jurisdiction scopes. Therefore, a spatiotemporal multi-granularity risk prediction, which enables adjustable predictive horizons and multiple spatial scales for risk predictions, is sponta-

• *Z. Zhou, Y. Wang*, X. Xie, L. Chen and C. Zhu are with University of Science and Technology of China, Suzhou, China.*
E-mail: {zzy0929, cll006, cczhu}@mail.ustc.edu.cn, {angyan,xkxie}@ustc.edu.cn*

Manuscript received May 7, 2020; revised xx xx, 2020.

neously required for satisfying the diversified requirements of transportation services, ranging from sequential urban route planning to multiple spatial-level traffic controlling. Recent pioneering multi-step prediction methods have been widely used in the field of traffic. For instance, [21] employed attention-enhanced encoder-decoder mechanisms to capture temporal correlations in road speeds, and [10], [22] utilized hierarchical graph structures to recursively extract sequential dependencies in taxi demands and traffic flows. Despite the superiority of graph convolution and hierarchical structure have been demonstrated, these methods on predicting continuous elements cannot be directly applied to predict accidents due to the sporadic nature and less conspicuous temporal tendency of accidents. Typically, Figure 1 reports the newly observed time-varying region-wise correlations and differential associations among urban traffics and accidents [1] in both New York City (NYC) and Suzhou Industry Park (SIP). These kinds of spatiotemporal correlations and differential associations have never been considered in previous accident predictions, and may inherently reduce the performances of previous works. Furthermore, as shown in Figure 1(a), there only exist two accident records during one selected 10-min interval in NYC, so the prediction performances will deteriorate with the increase of time steps. Therefore, it is even challenging to achieve urban traffic risk predictions with a spatiotemporal multi-granularity perspective.

In this paper, we propose a novel deep learning network to foresee citywide accident risks in a spatiotemporal multi-granularity fashion, where multiple spatial scales and temporal steps are jointly predicted. Specifically, we first summarize the sparse spatiotemporal traffic-related information into two categories and correspondingly provide respective solutions. Then we explicitly model correlations between time-varying traffic statuses and accidents with a carefully designed Differential Time-varying Graph Convolutional Network (DT-GCN). Finally, we alleviate the sequential error accumulation by feeding step-wise contextual factors into the decoder and further boost the performance of multi-step discrete accident prediction by leveraging three-scale highly correlated forecasting tasks. The contributions of our work are summarized as follows.

- To our best knowledge, this is the first work targeting spatiotemporal multi-granularity urban traffic risk prediction where the sporadic event prediction is transferred into a learnable self-adaptive ranking task. It provides a paradigmatic DNN-based solution to spatiotemporal multi-granularity forecasting of sporadic events.
- We take an initial step to systematically deal with the spatiotemporal sparsity challenges according to their origins. Based on observations in short-term traffics and accidents, we provide a novel node-wise proximity measurement and signal-wise differential operation integrated DT-GCN, to extract the time-varying region-wise correlations among urban traffics and accidents, and further benefit GCN community.

1. Note here the differential associations indicate the correlations among the variation of traffic volumes within adjacent time intervals and subsequent accidents in same subregions.

TABLE 1: Summarization of traffic accident prediction

Time granularity	Single Step	Multiple steps
Long-term	[23], [2]	[12], [13]
Short-term	[14], [24], [25], [17], [26], [16], [19], [27]	**Our work**

- We devise a novel hierarchical learning structure, Context-Guided LSTM, to decode multi-step risks in three spatial scales. The step-wise context is injected into the decoder to learn region-context interactions and consequently guides the multi-scale learning with risk assignment and gathering layers.

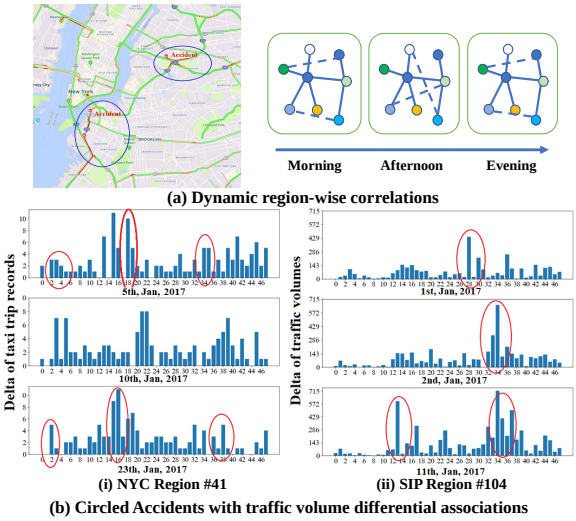


Fig. 1: Novel observations in the joint analysis of urban traffics and accidents. Subfigure (a) illustrates the correlations between congestion propagations and accident risks, the correlations between accident concurrences and similar road structures, as well as an the example of dynamic region-wise dependencies according to commute and tidal flows. Subfigure (b) illustrates the obvious differential associations between traffics and accidents, and here 'Delta' refers to the traffic volume variations within two adjacent intervals.

The rest of this paper is organized as follows. We first give preliminaries and formal definitions in Section 2. Then we detail our spatiotemporal multi-granularity accident forecasting in Section 3. Then extensive experiments and substantial ablation studies are conducted and demonstrated in Section 4. The related works are briefly reviewed in Section 5, followed up by further discussion in Section 6. Finally, we conclude our paper in Section 7.

2 Preliminaries and Definitions

In this section, we first present the preliminaries and some basic definitions of this paper, then formally define the problem studied in this paper.

In our work, we first divide the study area into q medium-sized rectangular regions ('rectangular regions' in short). Each rectangular region consists of several small-sized square subregions ('subregions' in short). The hierarchical division of NYC city is illustrated in Figure 2. We

assume there are totally m subregions in the study area, and subsequently model the m subregions with an urban graph.

Definition 1 (Urban Graph). *The study area can be defined as an undirected graph, called* Urban Graph $G(\mathcal{V}, \mathcal{E})$. *Here,* $\mathcal{V} = \{v_1, v_2, \cdots, v_m\}$, *where* v_i *denotes the* i*-th square-shaped urban subregion. Given two vertexes* $v_i, v_j \in \mathcal{V}$, *the edge* $e_{ij} \in \mathcal{E}$ *within these two vertexes indicates the connectedness between these two subregions, where*

$$e_{ij} = \begin{cases} 1 & \text{iff the traffic elements within two subregions have strong correlations} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Note that the traffic elements of a vertex consist of two aspects, static road network features and dynamic traffic features. **And for each subregion, we adaptively select the most ρ correlated nodes in the urban graph as its neighbors to reduce the computational complexity where ρ is the percentage of the selected neighbors versus the total nodes in the graph,** then the corresponding nonzero items in affinity matrix $\mathcal{A}_s$ and $\mathcal{A}_o^{\Delta t}$ (introduced in the next section) refer to the subregions with strong correlations.

The dynamic traffic features of subregion v_i in a specific time interval Δt can be modeled by l_d parts, e.g., (a) the intensity of human activities, represented by traffic volume $TV_{v_i}(\Delta t)$; (b) the traffic conditions, represented by the average traffic speed $a_{v_i}(\Delta t)$; and (c) the level of traffic accident risks $r_{v_i}(\Delta t)$. Formally, traffic features are defined as below.

Definition 2 (Static Road Network Features). *For urban subregion,* $v_i \in \mathcal{V}$, *the static features of road networks within the subregion, cover* l_s *statistical spatial attributes of the numbers of road lanes, road types, road segment lengths and widths, snow removal priorities and the numbers of overhead electronic signs, for all road segments inside, can be denoted as a fixed-length vector* s_i. *The static road network features of the entire urban region can be formulated as* $\mathcal{S} = \{s_1, s_2, \cdots, s_m\}$.

Definition 3 (Dynamic Traffic Features). *For* $v_i \in \mathcal{V}$, *the dynamic traffic features of* v_i *within a given interval* Δt *can be formulated as* $f_{v_i}(\Delta t) = \{TV_{v_i}(\Delta t), a_{v_i}(\Delta t), r_{v_i}(\Delta t)\}$. $r_{v_i}(\Delta t)$ *is the summation of the number of accidents weighted by the corresponding severity levels* [2]. *In particular,* $r_{v_i}(\Delta t) = \sum_{j=1}^{3} j * \tau_{v_i}^{\Delta t}(j)$, *where* j *indicates the type of accident severity,* $\tau_{v_i}^{\Delta t}(j)$ *denotes the number of accidents of type* j. *So the accident risk distributions and the dynamic traffic features of the entire urban domain within* Δt *can be represented by* $\mathcal{R}(\Delta t) = \{r_{v_1}(\Delta t), r_{v_2}(\Delta t), \cdots, r_{v_m}(\Delta t)\}$ *and* $\mathcal{F}(\Delta t) = \{f_{v_1}(\Delta t), f_{v_2}(\Delta t), \cdots, f_{v_m}(\Delta t)\}$, *respectively.*

Definition 4 (Multi-granularity Spatiotemporal Traffic Accident Prediction). *Given static road network features* $\mathcal{S}$ *and the historical dynamic traffic features* $\mathcal{F}(\Delta t)\,(\Delta t = 1, 2, \cdots T)$, *our task is to predict both coarse-grained and fine-grained accident distributions* $\mathcal{O}_C(\Delta t')$ *and* $\mathcal{O}_F(\Delta t')$, *along with the selected* M *high-risk subregions* $\mathcal{V}_M(\Delta t')$, *where* $\Delta t' = T+1, T+2, \cdots T+r$ *and* r *denotes the length of the target spatiotemporal accident series to forecast.*

2. We define three accident risk types: minor accidents, injured accidents, and fatal accidents [25]. We assign weights 1, 2, and 3 to the three types, respectively.

TABLE 2: Description of Notation

Symbol	Description
m	Number of subregions in the urban graph
$\mathcal{V} = \{v_i\}$	Spatial urban graph node set of subregions
$\mathcal{E} \in \mathbb{R}^{m \times m}$	Edges between connected nodes
$\mathcal{A}_s \in \mathbb{R}^{m \times m}$	Static affinity matrix
$\mathcal{A}_o^{\Delta t} \in \mathbb{R}^{m \times m}$	Dynamic overall affinity matrix in Δt
$\mathcal{S} \in \mathbb{R}^{m \times l_s}$	Static road network features in subregions
$\mathcal{R}(\Delta t) \in \mathbb{R}^{m \times 1}$	Citywide fine-grained risks in Δt
$\mathcal{F}(\Delta t) \in \mathbb{R}^{m \times l_d}$	Citywide dynamic traffic features in Δt
$\mathcal{O}_F \in \mathbb{R}^{m \times r}$	Citywide fine-grained risks in predicted r intervals
$\mathcal{O}_C \in \mathbb{R}^{q \times r}$	Citywide coarse-grained risks in predicted r intervals
$\mathcal{V}_M \in \mathbb{R}^{M \times r}$	High-risk subregions in predicted r intervals

All the mathematical notations that will be used in this paper have been defined and listed in Table 2.

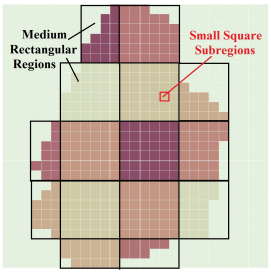


Fig. 2: Hierarchical division of NYC

3 Spatiotemporal Multi-Granularity Traffic Accident Forecasting

The spatiotemporal multi-granularity perspective in our task can be explained as predicting accidents for multiple time steps in both coarse-grained and fine-grained spatial granularities. In this section, we first show the overview of our proposed spatiotemporal multi-granularity traffic accident prediction framework RiskSeq, and elaborate its different modules.

3.1 Framework Overview

As illustrated in Figure 3, our proposed framework RiskSeq includes a data preprocessing component and two main modules: i) DT-GCN encoder module, and ii) Context-Guided LSTM decoder.

3.2 Data Preprocessing

Given the specific characteristics of traffic accidents, such as sparse and sporadic distribution, incomplete and heterogeneous multi-source information collection, we propose a series of strategies to jointly mitigate these issues.

3.2.1 Addressing Spatial Heterogeneities in Accident Prediction

As described in [27], high-risk regions tend to be focused on downtown, leading to spatial imbalance and neglect the

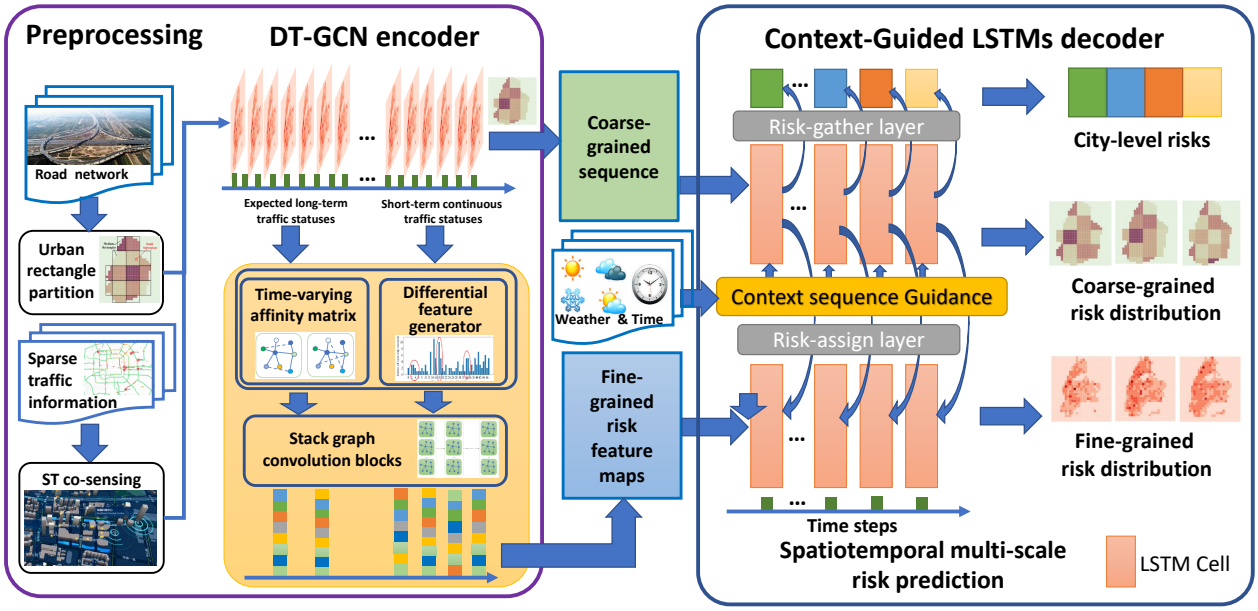


Fig. 3: Framework Overview of RiskSeq

relatively high-risk rural areas. Thus, in our work, the subregions are organized hierarchically as illustrated in Figure 2. The medium-sized rectangular regions and square-shaped subregions are responsible for collecting coarse-grained and fine-grained accident distributions, respectively. Then multiple spatial scales of distributions can be predicted and high-risk subregions in different medium-sized rectangular regions are highlighted with considering local risk statuses, especially benefiting urban areas in periphery.

3.2.2 Tackling Dual-sparsity Challenges in Accident Prediction

According to sparsity origins of sparse datasets, we categorize the sparsity information into two scenarios, *intrinsic sparsity* and *fake sparsity*. Regarding intrinsic sparsity, the sensed data is sparse and sporadic distributed due to the inherent sparse nature of itself. For instance, given interval Δt, there are seldom traffic accidents and most items in $\mathcal{R}(\Delta t)$ are zeros. For fake sparsity, the sensed data is spontaneously intensive, and this kind of sparsity is caused by the sparse distribution of sensing devices. For example, the traffic flows of road intersections captured by stationary surveillance cameras are fictitiously sparse due to the sparse sensor deployments. We demonstrate these two cases in Figure 4. Given the sparse nature of these two kinds of data, directly applying machine learning including deep learning methods will fall into zero-inflated issue [28], which predicts all results as zero values.

Overcoming zero-inflated issue in intrinsic sparsity issue. Deep Neural Networks (DNNs) suffer from zero-inflated issues and predict invalid results if the nonzero items in training labels are extremely rare [26], [28]. To discriminate a large number of zero risk values in short-term intervals and enhance the training feasibility, a priori knowledge-based data enhancement (PKDE) strategy is proposed. Specifically, for interval Δt, we transform zero items in risk sets $\mathcal{R}(\Delta t)$ to negative values that are different from each other and discriminated by their subregion-level statistical accident records.

Specifically, we replace zero-value risk of v_i in each time interval with the negative statistical accident intensity π_{v_i}:

$$\pi_{v_i} = b_1 \log_2 \varepsilon_{v_i} + b_2 \tag{2}$$

where ε_{v_i} is the statistical accident indicator quantifying the accident frequency of v_i among all subregions. b_1 and b_2 are the coefficients to maintain symmetry between the range of the absolute value of π_{v_i} and true risk values. It reflects the fact that a zero-item subregion is with lower accident risk than subregions with accidents, and the subregion with lower accident risk indicator has a lower accident probability, preserving the ranks of actual accident risks. The transformation ensures the accident intensity value negative and different from each other, enlarging the gap between the positive and negative samples.

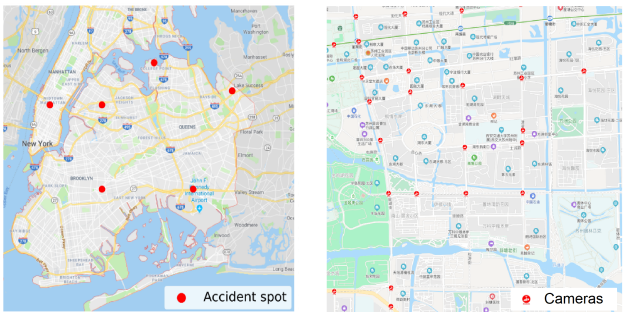
Complementing sparse sensing data in fake sparsity issue. The collected real-time traffic information for accident prediction is usually insufficient [29]. Fortunately, the dynamic traffic statuses tend to have interactive effects with the spatial road network structures [30]. We thus adopt a co-sensing strategy based on spatiotemporal deep factorization machine (ST-DFM) [31], by taking advantage of the static and contextual information.

The road network similarities between subregions are first extracted by static affinity matrix $\mathcal{A}_s$ where the item $\alpha_s(i,j)$ in $\mathcal{A}_s$ denotes static affinity between subregion v_i and v_j. The static affinity can be calculated by

$$\alpha_s(i,j) = \begin{cases} 1 & \text{if subregion } v_i \text{ and } v_j \text{ are adjacent} \\ e^{-JS(s_i \| s_j)} & \text{otherwise} \end{cases} \tag{3}$$

Here, the JS function is the Jensen-Shannon divergence [32] which measures the similarity between two distributions.

ST-DFM contains the Compressed Interaction Network (CIN) module and the DNN module. Multi-source features within three spatiotemporal fields i.e. static spatial features, dynamic traffic features and timestamps are embedded into a fixed-length vector. The CIN module learns the field-wise interactions in a vector-wise level while the DNN module projects features into high-level representations and finally obtains complex feature combinations. We infer speed values by feeding traffic volumes and corresponding static and contextual information at the same subregion into ST-DFM and vice versa. Therefore, the ST-DFM can be trained with the intersections of two real-time traffic datasets within the same spatiotemporal scopes and the citywide traffic information can thus be maximumly inferred.



(a) Accident spots in 10-min interval in NYC (b) Surveilliance camera distribution in SIP

Fig. 4: Illustration of intrinsic sparsity and fake sparsity. (a) There only exists 6 accidents in one 10-min interval of Jan, 01, 2017 in NYC, indicating the intrinsic sparsity of events like accidents. (b) There are only 23 deployed cameras in approximate 11 km^2 but flows are everywhere, indicating a fake sparsity.

3.3 DT-GCN based Spatiotemporal Encoder

In this section, we elaborate our proposed Differential Time-varying Graph neural Network (DT-GCN). As shown on the left part of Figure 1(a), the occurrences of accidents force the traffic flows to accumulate, eventually leading to the risk propagation along adjacent road segments. And subregions share similar both static (e.g. intersection structures) and dynamic traffic patterns may suffer accident concurrences with the same weather during near intervals. The core idea of GCN is to aggregate adjacent information and obtain local patterns with the designed aggregation matrix and learnable convolution kernels. Therefore, we inherit GCN as the basic framework in DT-GCN spatiotemporal encoder, taking advantage of its potential in modeling non-Euclidean correlations and subregion-wise risk propagations [30]. Here, we further propose the time-varying overall affinity and differential association generator as the aggregation matrix and novel graph signal operation by identifying distinct observations in very short-term accident datasets.

3.3.1 Time-varying Overall Affinity Matrix

Since it is difficult to capture accident patterns directly, we introduce general traffic statuses such as speeds, flows to help the prediction [33], [34]. Intuitively, the traffic statuses reveal spatial dependencies among each subregion [35], [36] and this kind of dependency is recently verified to be time-varying [36], [37], as described in Figure 1(a). Besides, a visualization of the backward differences of taxi trips/traffic flows on two datasets is shown in Figure 1(b). The undulant changes of traffic flows also provide the evidence for the necessity to model the time-varying correlations. Therefore, to specifically address our spatiotemporal multi-granularity predictions, we propose a time-varying overall affinity matrix $\mathcal{A}_o$ for measuring and aggregating the inter-subregion time-varying proximities. The time-varying overall affinity is calculated by three perspectives, (i) affinity of road network features, (ii) affinity of dynamic traffic statuses and (iii) transitions of traffic flows between subregions, where the former one is responsible for static similarity extraction while the latter two are responsible for capturing the dynamic spatial correlations. In interval Δt, the item $\alpha_o^{\Delta t}(i,j)$ in $\mathcal{A}_o^{\Delta t}$ denotes the dynamic overall affinity within subregions v_i and v_j:

$$\alpha_o^{\Delta t}(i,j) = e^{-JS(s_i^* \| s_j^*)} + \gamma * e^{-JS(C_i^{\Delta t} \| C_j^{\Delta t})} + \beta * tr_{ij}^{\Delta t} \quad (4)$$

$C_i^{\Delta t}$ includes the traffic volume $TV_{v_i}(\Delta t)$ and average speed $a_{v_i}(\Delta t)$ of subregion v_i within the same interval Δt in each day of last week. The $tr_{ij}^{\Delta t}$ is the element in matrix $TR^{\Delta t}$ and describes the average transitions of the interval Δt during last week. Notice that we modify the weights of static spatial attributes of subregions based on their different effects on accidents with an attention-based scheme. Also, the accident-based static features of subregion v_i can be denoted as s_i^*. Further, a weighted factor γ, β are used to adjust the proportion that each traffic proximity measurement accounts for the overall affinity. With such overall affinity, distant subregions but have potential accident-related correlations regarding traffic characteristics can also be connected dynamically. To transform the affinity matrix into spectral domain and utilize the first-order approximation, we calculate the normalized adjacent matrix $\mathcal{A}_C^{\Delta t}$ with $\mathcal{A}_o^{\Delta t}$ [38]. First, we derive $\mathcal{B}^{\Delta t}$:

$$\mathcal{B}^{\Delta t} = \mathcal{A}_o^{\Delta t} + I_m \quad (5)$$

where I_m is the identity matrix of $m \times m$. Second, we calculate $\Phi^{\Delta t}$ by

$$\Phi^{\Delta t} = \begin{bmatrix} \varphi_{11} & 0 & \cdots & 0 \\ 0 & \varphi_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \varphi_{mm} \end{bmatrix} \quad (6)$$

where $\varphi_{ii} = \sum_{j=1}^{m} b_{ij}$ and b_{ij} is the element in matrix $\mathcal{B}^{\Delta t}$. Then, we can obtain time-varying affinity-based normalized adjacent matrix for aggregation by

$$\mathcal{A}_C^{\Delta t} = \left(\Phi^{\Delta t}\right)^{-\frac{1}{2}} \mathcal{B}^{\Delta t} \left(\Phi^{\Delta t}\right)^{-\frac{1}{2}} \quad (7)$$

3.3.2 Differential GCN for Extracting Spatiotemporal Features

It is observed that accidents or events in the road network are more relevant to abnormal variations of urban traffic conditions [33], [39]. The intuition can be explained that the larger variations of fundamental traffic elements indicate the abnormal changes in the road network, thus increasing the possibility of accident occurrences. Fortunately, most cases

of accidents in Figure 1(b) verify the correctness of this intuition. To this end, we introduce a differential association generator to calculate differential images within adjacent time intervals. By feeding the differential images into GCN, the propagations and interactions of abnormal changes in traffic can be modeled and the immediate correlations between traffic condition variations and accidents are learned. Given Δt, the differential vector $\overrightarrow{\Theta}^{\Delta t}$ can be computed by

$$\overrightarrow{\Theta}^{\Delta t} = \mathcal{D}(\Delta t) - \mathcal{D}(\Delta t - 1) \tag{8}$$

where $\mathcal{D}(\Delta t) = \{d_{v_1}(\Delta t), d_{v_2}(\Delta t), \cdots, d_{v_m}(\Delta t)\}$ and $d_{v_i}(\Delta t) = \{TV_{v_i}(\Delta t), a_{v_i}(\Delta t)\}$. For all subregions in Δt, by combining their dynamic traffic features and the corresponding differential vectors, we generate a united feature tuple $\mathcal{U}(\Delta t) = \left\{\mathcal{F}(\Delta t), \overrightarrow{\Theta}^{\Delta t}\right\}$.

3.3.3 Long-term and Short-term Encoders

Traffic statuses and risks in subsequent time steps are determined by both long-term expectations like seasonal influences and short-term instantaneous statuses such as recent trends and unexpected incidents [22]. Here, we separately encode long-term expectations and short-term statuses. Specifically, as illustrated in Figure 3, given Δt, we first retrieve the fine-grained united feature tuples $\mathcal{U}(\cdot)$ for the same interval Δt in the last τ weeks and the recent τ days respectively, and denote them as the weekly and daily components of the training sample. Next, the average values of observations are calculated for both the corresponding weekly and daily components of this sample, and are taken as two distinctive inputs of the long-term DT-GCN encoder. After then, we take the most recent h time intervals as the short-term instantaneous traffic inputs of our DT-GCN encoder [3]. The detailed architecture of one individual DT-GCN is demonstrated in Figure 5, where $\oplus$ denotes element-wise addition in residual shortcut connections [40]. For interval Δt, we denote the corresponding feature tuple set as $\mathbb{U}_*^{\Delta t}$. The GCN works recursively as,

$$\mathcal{H}_{n+1} = \text{Leaky\_ReLU}(\mathcal{A}_C^{\Delta t}\mathcal{H}_n\mathcal{W}_n) \text{ where } \mathcal{H}_0 = \mathbb{U}_*^{\Delta t} \tag{9}$$

Here $\mathcal{H}_n$ and $\mathcal{W}_n$ indicate the hidden representations of the nth layer graph convolution block and the weights of the corresponding convolution kernels, respectively. The learnable kernels can automatically distinguish the importances of region-wise correlations and aggregate adjacent graph representations from three different perspectives of the time-varying overall affinity. By employing several residual connections, we also combine the low-level convolution feature maps with the high-level feature maps to capture the multi-hop node-wise correlations, and subsequently enhance the graph representation [41]. Noted that the Batch Normalization (BN) operations are inserted into every 2 GCN layers to avoid gradient explosions. Considering the negative values in the dataset we transformed, we select the Leaky_ReLU as the activation function. In addition, the contextual external factors, i.e., timestamps and meteorological data, are embedded into a fixed-length vector consecutively, and then are fused with the outputs of GCN blocks. Finally, for each

3. According to the settings in [10], we here set the values of τ and h as 3 and 6, respectively.

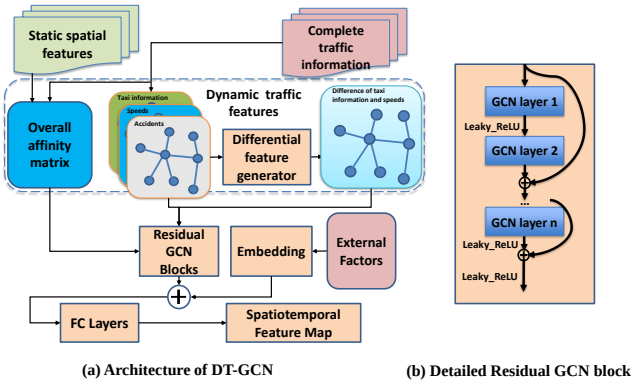


Fig. 5: Architecture of one individual DT-GCN

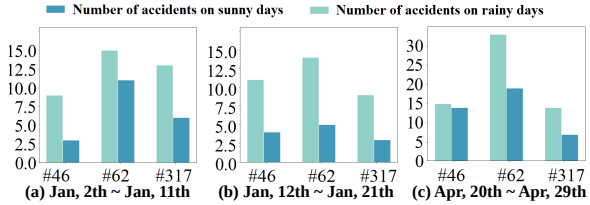


Fig. 6: Accident statistics under different contexts. It summarizes the accident occurrences of three subregions #46, #62, #317 in NYC during three selected 10-day periods in 2017 which all contain 5 rainy days and 5 days without rain.

interval, the output of the DT-GCN encoder is mapped into a one-channel feature map. Given the sequential nature of the long-term and short-term inputs, the outputs of the DT-GCN encoder are consequently formulated as a fine-grained risk-map sequence $\mathcal{M}_F = \{\mathcal{M}_F{}^0, \mathcal{M}_F{}^1, ..., \mathcal{M}_F{}^{h+1}\}$.

3.4 CG-LSTM based Spatiotemporal Decoder

Extensive experiments reveal that, due to the possible severe error accumulation of RNN-based methods [10], [42], the forecasting performance declines rapidly with the increase of prediction steps. Based on the analysis of real-world historical data, we discover that some time-sensitive contextual factors such as meteorology can significantly influence the occurrences of traffic accidents, especially in some specific subregions or road segments where the traffic volumes are relatively stable with different weathers. For instance, in SIP, there are 2.26 accidents occurring averagely on rainy days while there are only 1.89 accidents averagely on one sunny days. Figure 6 illustrats some cases of the correlations among the contextual factors and accidents in NYC. To this end, both the spatiotemporal correlations and the contextual factors should be carefully involved in step-wise future traffic accident predictions.

Based on real-world data analysis, we further discover that urban accidents usually follow clustering distributions in spatial perspective, and local accident risks can be influenced by surrounding traffic statuses. Thus, considering above two factors, we design a novel CG-LSTM decoder, which employs a hierarchical sequential learning structure to jointly learn accident distributions in both coarse-grained

and fine-grained granularities with contextual factors involved, to achieve high-quality accident predictions.

The core idea of the CG-LSTM decoder is to guide the learning of step-wise accident maps with contextual factors, and to enhance the spatial representations in the hierarchical LSTMs with multiple tasks. CG-LSTM decoder consists of two parallel LSTM components: F-LSTM and C-LSTM. First, the sequential outputs of the DT-GCN encoder and the aggregated spatial coarse-grained risk maps are considered as two separate inputs for F-LSTM and C-LSTM. Given the intensive nature of coarse-grained risk distributions, we then take C-LSTM as an intermediate to enhance a series of urban graph representations. Specifically, C-LSTM sequentially receives the contextual factors of each future time step and adds them into the hidden states learned in the previous step. The combined hidden vector with contextual factors then represents the potential importance of contextual factors specific to each subregion. Given interval Δt, we denote the coarse-grained risk map and hidden state in the C-LSTM Cell as $\mathcal{M}_C{}^{\Delta t}$ and $\mathcal{I}_C{}^{\Delta t}$ respectively, and the hidden state in $\Delta t+1$ can be updated by previous states and current context factors:

$$\mathcal{I}_C^{\Delta t+1} = \mathrm{LSTM_C}(\mathcal{M}_C^{\Delta t+1}, [W_{\mathrm{ext}} * E^{\Delta t+1} + \mathcal{I}_C^{\Delta t}]) \quad (10)$$

where $E^{\Delta t}$ represents the context-guided factors, and W_{ext} refers to the context alignment weights which are used to adapt the same dimension with $\mathcal{I}_C^{\Delta t}$. So far, the C-LSTM structure can easily capture the step-wise accident-context interactions, and can adaptively control the risks and mitigate the error accumulation in sequential predictions. To further guide the learning process, we design a risk-assign layer to propagate the context influences to fine-grained distributions in F-LSTM by learning the latent hierarchical spatial correlations. Then the learned risk assignments are added into the previous-step hidden representations element-wisely for subsequent temporal dependency learning. Regarding the fine-grained risk learning in F-LSTM, the risks can be learned from two aspects, the spatial backbones of risk distributions obtained from DT-GCN and the risk intensities controlled by the hidden representations in C-LSTM, hence the risk representations can adaptively learn both self and neighborhood dependencies with temporal modeling. Given Δt, the learned hidden states $\mathcal{I}_F{}^{\Delta t+1}$ in F-LSTM can be modified by:

$$\mathcal{I}_F^{\Delta t+1} = \mathrm{LSTM_F}(\mathcal{M}_F^{\Delta t+1}, [W_{\mathrm{asgn}} * \mathcal{I}_C^{\Delta t} + \mathcal{I}_F^{\Delta t}]) \quad (11)$$

where $W_{\mathrm{asgn}} \in \mathbb{R}^{I_f \times I_c}$ indicates the learnable weights in the risk-assign layer, I_c and I_f represent the hidden dimensions in C-LSTM and F-LSTM respectively.

Similarly, the counterpart risk-gather layer performs graph-coarsen operations to gather coarse-grained risks into a graph-level summation of accident records $\tilde{R}_S^{\Delta t}$, namely, the city-level risk indicator of interval Δt.

$$\tilde{R}_S^{\Delta t} = W_{gath} * \mathcal{I}_C^{\Delta t} \quad (12)$$

where $W_{gath} \in \mathbb{R}^{1 \times I_c}$ indicates the learnable weights of the risk-gather layer. The hidden states in both F-LSTM and C-LSTM will be further mapped into the same dimension with their corresponding input sequence. We eventually obtain the learned spatiotemporal multi-granularity risks by:

$$\mathcal{O}_F^{\Delta t} = \mathrm{Leaky\_ReLu}(W_{RF} * \mathcal{I}_F^{\Delta t} + \mathrm{b}_{RF}) \quad (13)$$

$$\mathcal{O}_C^{\Delta t} = \mathrm{ReLu}(W_{CF} * \mathcal{I}_C^{\Delta t} + \mathrm{b}_{CF}) \quad (14)$$

Since the fine-grained risk labels are partially negative, and the coarse-grained risks are all positive, **we adopt** Leaky_ReLU **and** ReLU **as their activation functions, respectively. Here,** $W_{RF} \in \mathbb{R}^{m \times I_f}$ **and** $b_{RF} \in \mathbb{R}^{m \times 1}$ **are the weights and biases for mapping layers of fine-grained risks while** $W_{CF} \in \mathbb{R}^{q \times I_c}$**,** $b_{CF} \in \mathbb{R}^{q \times 1}$ **are the weights and biases for layers aggregating coarse-grained risks,** $\mathcal{O}_F^{\Delta t}$ **and** $\mathcal{O}_C^{\Delta t}$ **are the learned fine-grained and coarse-grained risk distributions respectively.** The three spatial scales of accident risk learning can not only be viewed as multi-granularity predictions, but also can jointly optimize representation abilities as a task-wise regularization.

3.5 Most-likely Accident Region Selection

For selecting the most-likely accident subregions, we devise an adaptive high-risk region selection mechanism with considering both the spatial heterogeneity issue and time-varying citywide risk levels. Specifically, the risk-assign connections between the multi-scale spatial risk distributions allow the fine-grained risks to take peripheral urban areas into account and adequately absorb the hierarchical correlations. For Δt, we take the learned summational risks $\tilde{R}_S^{\Delta t}$ as the citywide risk indicator and let the adaptive threshold of the high-risk subregion number be $K(\Delta t)$ equalling to $\tilde{R}_S^{\Delta t}$. Regarding each interval, we select $K(\Delta t)$ subregions with the highest risks from $\mathcal{O}_F^{\Delta t}$ as a set of most-likely accident subregions $\mathcal{V}_M$. Then, the learned $K(\Delta t)$ reduces the number of over-predicted regions and keeps the outputs conform to the time-sensitive changes of contextual factors.

3.6 Optimization

The r tuple outputs $\{< \mathcal{O}_F^{T+1}, \mathcal{O}_C^{T+1}, \tilde{R}_S^{T+1} >, ..., < \mathcal{O}_F^{T+r}, \mathcal{O}_C^{T+r}, \tilde{R}_S^{T+r} >\}$ constitute a predicted spatiotemporal multi-granularity accident risk sequence, where each tuple denotes the results of one time step. In the training process, we have the total loss of this multi-task risk-oriented learning framework:

$$\mathrm{Loss}(\theta) = MSE_F + \lambda_1 * MSE_C + \lambda_2 * MSE_R + \lambda_3 * \mathrm{L2} \quad (15)$$

where θ represents all learnable parameters in our framework. MSE_F, MSE_C and MSE_R are the mean square errors of the risks in fine-grained, coarse-grained, and citywide scales. We here employ L2 regularization to avoid the overfitting issue, and use λ_1, λ_2, λ_3 as the hyperparameters of the loss function.

For optimizing the algorithm, we introduce Adam optimizer [43]. The learning rate, which has a decay of 0.98 in every 10 epochs, is initialized as 0.001. Early stopping technique is also applied during the training process to avoid overfitting.

4 Experimental Studies

In this section, we conduct extensive experiments to evaluate our method for spatiotemporal multi-granularity traffic accident prediction from multiple perspectives, including performance comparisons, ablation studies and case studies.

4.1 Data Description

The experiments are conducted on two real-world datasets: NYC Opendata between 1st Jan, 2017 and 31st May, 2017, and Suzhou Industrial Park (SIP) dataset between 1st Jan, 2017 and 31st March, 2017. For NYC dataset, we utilize the taxi trip volumes in each subregion as the indicator of human mobilities. For SIP dataset, it only contains traffic flows and speeds and we integrate it with another traffic accident dataset collected from Microblog, Sina, a social media platform. The statistics are shown in Table 3.

TABLE 3: Datasets statistics

City	Dataset[5]	Time Span	# of Regions	# of Records
NYC	Accidents	01/01/2017-05/31/2017	354	254k
	Taxi Trips			48,496k
	Speed Values			125k
	Weathers			604
	Demographics	Investigated in 2016		195
	Road Network			102k
SIP	Accidents	01/01/2017-03/31/2017	108	183
	Traffic Flows			1,399k
	Speed Values			311k
	Weathers			180

4.2 Experimental Settings

4.2.1 Implementation Details

In our experiments, we select 60%, 30% and 10% of dataset for training, evaluation and validation, respectively. The whole city of NYC is partitioned by small squares sized 1.5 km $\times$1.5 km and obtain 354 square-shaped subregions and 18 rectangular regions. In SIP dataset, 108 surveillance spots are gathered into 6 rectangular regions [4]. The accidents are transferred into corresponding two-scale risk distributions. All default settings of parameters involved in our framework are summarized in Table 4. Then the missing values and zero-value risks are complementing with ST-DFM and PKDE strategies to enhance the performance. Due to the incomplete accident records on Microblog, we omit the input of accidents in our framework and maintain the main components for traffic indications.

During training periods, dynamic traffic data and affinity matrices are aggregated into three groups, which consist of two expected historical observations and a sequence indicating short-term instaneous dynamics. The RiskSeq is trained with backpropagations and Adam method [43]. We eventually attain both coarse-grained and fine-grained accident distributions in the following 6 time steps and select the most-likely accident regions according to rankings.

TABLE 4: Parameter Settings during training period

Symbol	Description	Value
Δt	Granularity of time intervals	10 min
ρ	Connectedness of urban graph	10%
h	Length of recent risk sequence	6
r	Multi-step horizons	6
(I_f, I_c)	Hidden state dimensions in fine-grained and coarse-grained feature maps	(256, 48)
(γ, β)	Importances of elements in α_o	(1, 0.8)
$(\lambda_1, \lambda_2, \lambda_3)$	Weights of loss function	(1.2, 0.8, 1e-4)
-	Number of GCN blocks	6
-	Number of GCN learnable kernels	256

4.2.2 Evaluation Metrics

We evaluate our proposed RiskSeq from two perspectives [26]. (1) Regression perspective: Mean Square Error (MSE) of predicted risks. (2) Spatial classification perspective: a) Accuracy of top M (Acc@M) [44], which is widely applied in spatiotemporal ranking tasks, indicates the percentage of accurate predictions in subregions within M highest risks. Considering the actual capacity of urban traffic adminstration [45], we select approximate 5% subregions as the most-likely accident regions for comparison in our test. Thus, M equals 20 and 6 in NYC and SIP dataset respectively, that means subregions with 20 and 6 highest risks in NYC and SIP will be considered as high-risk subregions to compare with real-world accident records. b) Acc@K is an adaptive selection metric where K is the learned city-level risks in our framework.

4. The settings of the spatiotemporal partition should leverage the tradeoff between the accuracy and spatiotemporal granularity. Note that such a setting may be related to the results of accident prediction but is orthogonal to the generalities of our proposals.

5. It refers to different traffic-related records in the city.

4.2.3 Baselines

Eight competitive baselines for spatiotemporal prediction which have the potential to solve our task are as follows. **For fair comparison, we realize all these baselines to predict next 6 step accident risks with 12 previous time steps and three influential factors as ours (i.e. traffic volumes, average speeds and accidents) unless specified. All the hyperparameter settings of baselines are initialized based on their literatures and codes, and then we fine-tune them on our dataset and make themselves achieve their optimal performances.**

(1) ARIMA is a classic machine learning algorithm, well-known for predicting future values, especially for time series. **Here we utilize the accident time series and the parameter tuple in ARIMA (p, d, q) is set as $(1, 2, 6)$.**

(2) LSTM is a classic deep learning-based time series modeling method with long short-term memory module [16]. **We realize this LSTM with 64 neurons in each hidden layer.**

(3) Hetero-ConvLSTM is an advanced deep learning framework for traffic accident prediction [13]. The sizes of maps of NYC and SIP are 27$\times$27 and 15$\times$10, and the convolution kernels are set as 3$\times$3 for both two tasks. We involve previous 6 time steps to predict the next 6-step risks.

(4) STGCN is a multi-step traffic forecasting model, integrating graph convolution and gated temporal convolution by several spatiotemporal convolutional blocks [46]. We realize it by stacking two ST-Conv blocks, and each block consists of three layers with 64, 64, 64 filters.

(5) STG2Seq targets multi-step citywide passenger demand prediction based on an urban graph, and it employs a hierarchical graph convolutional structure to capture both spatial and temporal correlations simultaneously [10]. We

set 6 GCN blocks with 32 filters, and set the sizes of both sliding window and patch as 3.

(6) STSGCN is designed for spatiotemporal data forecasting, which captures localized spatiotemporal correlations and heterogeneities with a synchronous network [11]. We incorporate 4 synchronous graph convolutional layers and each layer includes 3 convolution operations with 64, 64, 64 filters.

(7) STDN proposes the flow gating and shifted attention to jointly model volume and flow interactions, and to address temporal shifting issues in spatiotemporal forecasting [7]. We stack 3 CNN layers and each convolution kernel is 3×3 with 64 filters. Each neighborhood accounts for $7\times$ 7 grids and hidden dimensions of each LSTM is set as 128.

(8) DFN combines a hierarchical recurrent structure with a context-aware embedding module to perform daily accident prediction [12]. The spatial embedding size, hidden layer and attention dimensions are all set to 32.

(9) MTPSO is a turbulent PSO-based method targeting time-series prediction, which introduces fuzzy relationships for robust predictions [3]. We realize it with three groups of influential factors and 12 previous time steps.

4.3 Evaluation Results

4.3.1 Comparison Performance

The comprehensive performances are illustrated in Table 5, which are the averaged errors and accuracies on all time steps. MSE-F and Acc@M measure the performance of fine-grained forecasting while MSE-C evaluates the coarse-grained prediction. We sum up the corresponding fine-grained risks to coarse-grained ones for baseline methods as they lack this output.

Encouragingly, RiskSeq achieves the highest accuracies and low MSE among all compared methods. On NYC dataset, our solution improves the best baseline by more than 4% on Acc@20. With limited 180 events in SIP, our RiskSeq achieves the highest accuracy of 71.27%, which surpasses the best baseline by nearly 5%. It means that more than 55% and 70% of real-world accidents are hit by our model on top-20 in NYC and top-6 in SIP, respectively. The reasons for relatively higher performance in SIP may lie in that the events are few and also regularly occur in the limited intersections. It is not astonishing that MSE-C values are consistently larger than MSE-F, as the coarse risks are risk summations of square subregions, but it still reveals the superiority of our multi-task learning that this scheme can enhance the multi-scale risk representations and improve the performance. The acceptable coarse-grained results (MSE-C) can reflect the effectiveness of RiskSeq in high-level coarsen risk modeling and the scalability of our method in smaller or medium-sized cities.

ARIMA and MTPSO take temporal dependencies into account while deep learning models can simultaneously encode both spatial and temporal correlations, and reasonably most deep models perform better than ARIMA and MTPSO. Thanks to the separated long-term and short-term modeling as well as the graph convolutions and gated mechanisms, STGCN and STG2Seq are capable of capturing short-term traffic variations and achieve better results than other baselines. Even though Hetero-ConvLSTM and DFN which focus on daily predictions try to consider the spatial heterogeneities with ConvLSTM blocks ensembled and multi-scale temporal dependencies with hierarchically structured recurrent framework, respectively, they still cannot adapt the multi-step short-term event forecasting due to irregular-shaped urban areas and extremely sporadic event distribution. Noticed that the state-of-the-art method STSGCN performs worst among all deep methods, and this may be ascribed to the rescaled adjacent matrices and redundant connections between adjacent intervals. Moreover, STSGCN, **STDN** and Hetero-ConvLSTM are with high computation costs for their ensembles, pixel-wise operations and extended adjacent matrix. Also, all baselines fail to consider the available time-sensitive influences and hierarchical spatial dependencies into step-wise forecasting, hence they may lack the capability of predicting multi-step events.

TABLE 5: Comprehensive performance comparisons

	NYC/SIP		
Models	Acc@20/Acc@6	MSE-F	MSE-C
ARIMA	20.72/30.63	0.0192/ 0.0162	0.0492/0.2215
LSTM	28.98/35.70	0.0179/0.0255	0.0477/0.2694
Hetero-ConvLSTM	28.03/34.84	0.0161/0.0487	0.1015/0.4039
STGCN	50.42/51.27	0.0188 /0.0452	0.0492/0.2885
STG2Seq	52.08/54.30	0.0138/0.0364	0.0693/0.1667
STSGCN	26.46/33.59	0.0183/0.0236	0.1285/0.3473
STDN	37.48/42.18	0.0203 /0.0354	0.0853/0.2142
DFN	40.26/36.98	0.0194 /0.0376	0.0548/0.2278
MTPSO	30.81/33.69	0.0218 /0.0420	0.0393/0.2065
RiskSeq	**56.42/71.27**	**0.0158/0.0401**	**0.0443/0.2702**

4.3.2 Evaluations on Stability of Multi-step Performance

To evaluate the long-term stability of our solution, we illustrate the step-wise accuracy in the following 6 time steps among all methods in Figure 7.

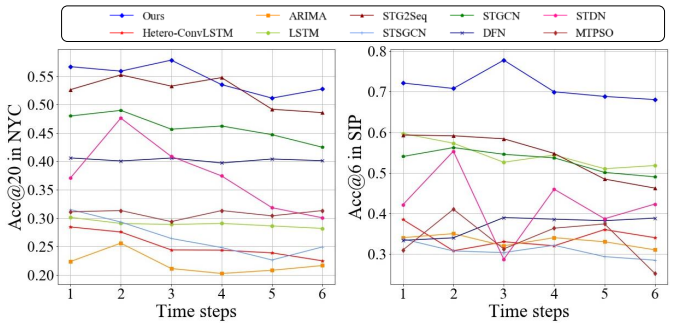
Intuitively, STG2Seq, STGCN and our RiskSeq perform much better than others due to their nice property in GCN-based sequential modeling. Specifically, we observe that the performance of our method keeps steady, retaining highest accuracy of more than 50% and 65% even at the last time step in NYC and SIP. **An interesting finding comes that RiskSeq achieves the best performance at the third step which may be attributed to the nature of traffic variations and the selection of time steps in spatiotemporal encoders.** By these results, the potential superiority of the combination of LSTM and hierarchical context-guided mechanism for capturing both contextual interactions and spatiotemporal dependencies is practically confirmed.

4.4 Ablation Study

To evaluate the importance of each proposed component in addressing challenging issues, we perform the ablation studies from two perspectives, i.e., dual-sparsity challenges, and spatiotemporal dependency modeling.

4.4.1 Dual-sparsity Challenges

As discussed above, spatiotemporal data mining usually suffers two categories of sparsity, i.e., fake and intrinsic. To verify whether the proposed data preprocessing method makes sense, we omit the PKDE data augmentation and



(a) Step-wise performance in NYC (b) Step-wise performance in SIP

Fig. 7: Evualtion on step-wise performance in NYC and SIP

co-sensing network separately with remaining components named RS-PKDE and RS-DFM. In Table 6, without data enhancement and co-sensing strategy, the performances suffer a sharp decrease of 37.86% and 13.37% in NYC and also show an obvious downtrend to 35.48% and 58.94% in SIP, verifying the effectiveness of our well-designed strategies.

4.4.2 Spatiotemporal Dependency Modeling

In spatiotemporal modeling views, we subsequently remove or replace some components as the ablative variants.

(1) RS-OA: We replace the **time-varying Overall Affinity** with the static adjacent matrix which calculated by the static similarities based on Eq (3).

(2) RS-DG: We remove **Differential association Generator** directly in the ablative version.

(3) RS-RC: We cut off the **Residual Connections** in DT-GCN in this variant.

(4) RS-CF: We omit the **Contextual Factor inputs** for guiding the decoder learning and let it learn without step-wise time-sensitive contexts.

(5) RS-CGLSTM: We remove the inputs of coarse-grained maps and replace the **CG-LSTM** with only one LSTM as the sequence decoder.

As can be seen, the integrated RiskSeq outperforms all its ablative variants on both datasets. We observe that the time-varying overall affinity contributes to the most remarkable improvement which is up to 18% in NYC and 4% in SIP on accuracy metric. With Residual Connections and Differential feature Generator in DT-GCN, our framework is able of aggregating mixed high-order correlations among subregions and capturing the abnormal changes within short terms, resulting in the improvements ranging from 1.82% to 13.63%. In addition, by removing the context-guided mechanism, we obtain the average results of 43.04%. The performance of RS-CGLSTM is most aprroximate to the results of integrated RiskSeq with the gap of 8%. It implicates the Occam's Razora principle that a light-weight model may perform better than the average. From the exciting results, we conclude that all well-designed components in RiskSeq exactly play important roles in our spatiotemporal multi-granularity prediction.

4.5 Hyperparameter Tuning

To illustrate how different hyperparameters affect the performance of the proposed framework, we show the tuning

TABLE 6: Ablative variants performance on two datasets

	NYC/SIP		
Variant	MSE	Acc@20(Acc@6)	Acc@K
RS-PKDE	0.0053/0.0512	18.56/35.48	16.28/29.45
RS-DFM	0.1260/0.0216	43.05/58.94	38.29/46.28
RS-OA	0.0116/0.0127	37.57/67.16	32.47/61.15
RS-DG	0.0118/0.0136	46.45/68.52	39.19/55.27
RS-RC	0.0208/0.0082	41.79/69.45	38.19/56.33
RS-CF	0.0123/0.0355	43.04/67.83	33.21/50.18
RS-CGLSTM	0.0128/0.0060	48.45/67.19	-
Integrated RS	**0.0158/0.0040**	**56.42/71.27**	**47.18/65.26**

TABLE 7: Performance on different train/test ratios

Ratio of Train/Test	2:1	3:1	4:1	5:1
Performance(Acc@20)	9.11	36.72	41.55	45.89
Ratio of Train/Test	6:1	7:1	8:1	9:1
Performance(Acc@20)	48.11	36.04	33.07	31.92

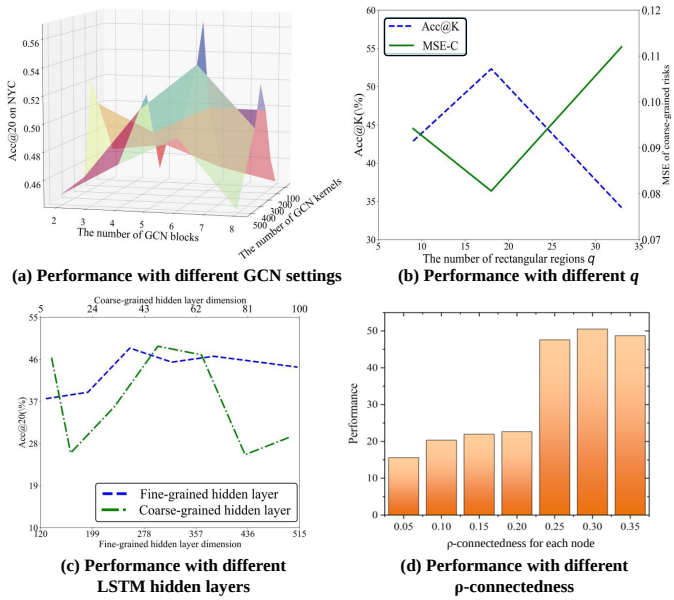
process of hyperparameters on NYC dataset. First, we adjust the number of GCN blocks and filters in each layer to make itself reach their best performance. It arrives the best performance at 6 GCN blocks and 256 kernels in each layer because the deep GCN layers should maintain an equilibrium between the robustness and algorithm complexity. And q equals 18 among $\{9, 18, 33\}$ when the Acc@20 arrives the highest at 53% approximately. Intuitively, the larger q induces less subregions included in one rectangular regions and vice versa. **By equilibrating the tradeoff between the serious zero-inflated issue in coarse-grained risk learning and the redundant and complex inter correlations, we finally obtain 18 rectangular subregions in NYC based on extensive experimental results.** We show the performance varying with the number of DT-GCN layers, filter kernels and different q in Figure 8.

For multi-task learning, we fix the weight of main task as 1, and tune λ_1, λ_2 by grid searching. Similarly, the searching is also performed on the importance of dynamic elements in overall affinity, and it reaches the best when (γ, β) equals $(1, 0.8)$. We summarize the performance comparison in terms of λ, γ, β, in Table 8.

We also investigate the influences of the ratio of the training samples versus the testing samples, and our RiskSeq performs better when the train-test ratio ranges from 5:1 to 6:1 in Table 7. It is consistent with the fact that more training samples can help capture deep spatiotemporal correlations and reduce epistemic uncertainty, but excessive data may also increase the aleatoric uncertainty and noises reversely.

TABLE 8: Performance on different hyperparameter settings

λ_1	λ_2	Acc@20	γ	β	Acc@20
0.8	0.8	38.16	0.5	0.5	41.02
0.8	1.2	41.14	0.5	0.8	49.61
1	1	45.71	0.8	0.5	46.51
1.2	0.8	54.26	0.8	0.8	49.63
1.2	1	53.28	0.8	1	51.40
1.5	0.8	47.19	1	0.5	51.50
1.5	1	48.74	1	0.8	52.65



(a) Performance with different GCN settings

(b) Performance with different q

(c) Performance with different LSTM hidden layers

(d) Performance with different ρ-connectedness

Fig. 8: Performance on different parameter settings

4.6 Efficiency of RiskSeq

We implement the proposed framework RiskSeq on a single Tesla V100 with 16GB. Python 3.6 and Tensorflow 1.9.0 libraries are involved to help build deep graph neural networks. The framework is trained offline and parameters learned are utilized for the online prediction. Here we present the analysis of time complexity of our framework. Let the number of neurons in each GCN block be O(n), the number of parameters in our DT-GCN is O(n^2). In our implementation, n is set to 256 and GCN contains 6 blocks. The total number of parameters is 256×256×6. In our testing, it takes an average of 5.6 seconds to do one round of accident forecasting, which sufficiently meets the requirement of real-time multi-granularity forecasting.

4.7 Case Study

To provide an intuitive understanding of our RiskSeq, we visualize the results in the following two scenarios.

The period-oriented evaluations are presented in Figure 9(a), and we also collect the corresponding precipitation [6]. As shown, risks are marked in a distinctive way and the highlighted subregions show spatial similarities with the ground truth. Manhattan District is always with higher risks and more accidents probably due to its highly dynamic traffic conditions and overloaded crowd flows. As observed, mornings and evenings tend to suffer fewer accidents while it becomes different in the afternoon. This is because fewer people will go for work on weekend mornings and they may go out for leisurely activities in the afternoon. The increasing number of vehicles in the road network and the identified rain subsequently lead to an accident-prone road situation around 2 p.m. In the evening, the color of risk map becomes deeper and risks are mostly focused in northern NYC, as the inherently higher risks in the night, and both nightclubs and bars are concentrated in Bronx District, the north of NYC. The results verify the motivation and effectiveness of time-varying region-wise modeling as well as the context-guided learning in different typical intervals.

Figure 9(b) shows the sequential results integrally derived by RiskSeq. At 10 a.m., the accidents are sporadically distributed in the city and there is only one accident in Manhattan subregion. However, with too many officers hurry to their working places, the accident circle, especially Manhattan subregion, is expanding and the sporadic accidents gradually evolve into three clusters. The reasons may boil down to the fluctuations of traffic volumes and the abnormal traffic changes around 10 a.m. Once the accidents occur in the crowded subregions, the accumulated vehicles tend to propagate the risks along the road segments from the accident spot centers. The rainy days make it worse. Furthermore, the aggregated accident clusters may follow a hierarchical distribution. Therefore, the competitive results demonstrate the potential superiority of propagation scheme and differential association structure in GCN as well as the hierarchical sequential learning in CG-LSTM decoder.

We conclude our careful-designed RiskSeq can not only perform well on sequential learning, but capture the time-varying dependencies during different typical periods.

5 Discussion

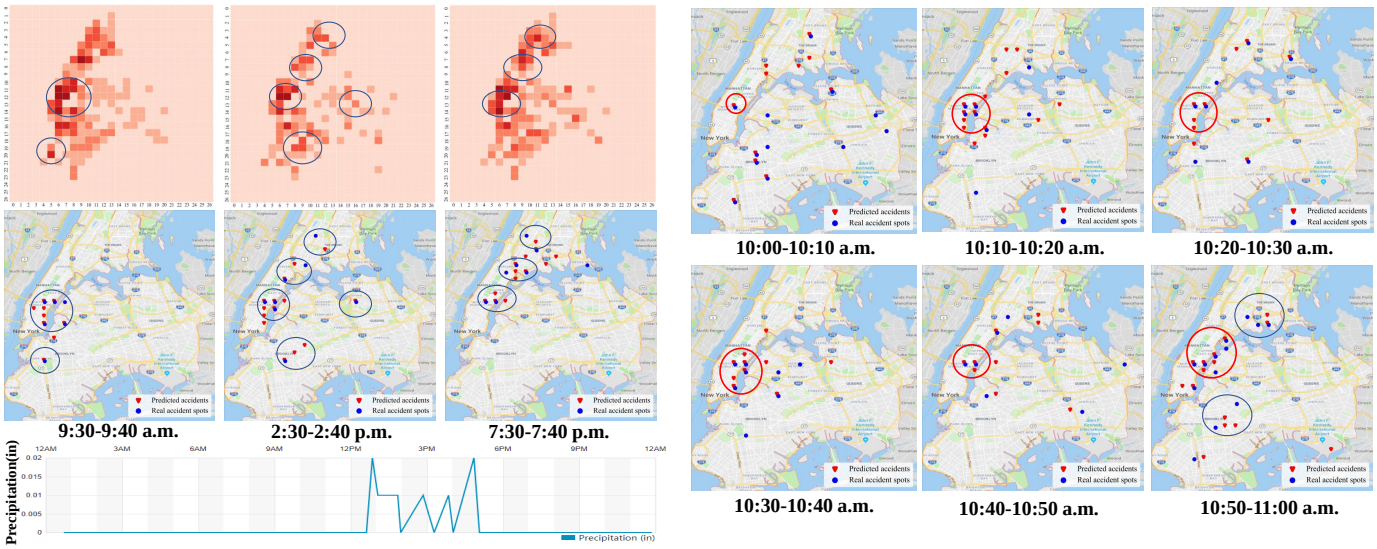
In this section, we discuss some related interesting issues.

General Applicability of RiskSeq. The core idea of RiskSeq is to dynamically aggregate neighborhood graph signals for better risk representations and to enhance interval-level predictions by employing step-wise context injections and multi-scale learnings. Besides promising performances of accident predictions, our work has the potential to benefit other downstream tasks in spatiotemporal forecasting. Crimes and epidemics share similar properties with traffic accidents, which occur occasionally and also exhibit interactions between spatial dependencies and human mobilities. After mitigating data incompletions with ST-DFM and urban covariates, and alleviating the issue of rare events with PKDE, multiple urban data sources are formulated into a graph. Human-related data as well as task-specific historical records are fed into DT-GCN for capturing time-varying and abnormal situations, and the multi-step predictions are boosted with CG-LSTM decoder.

Novel insights provided by RiskSeq. Targeting two inevitable sparse scenarios, we systematically address both intrinsic and fake sparsity by proposing novel strategies. We transfer the sparse event prediction into a learnable regression and ranking task which can be solved with DNN. This inspires researchers to mine the inherent and latent correlations in spatiotemporal sparse datasets from the perspective of sparsity origins. Novel sparsity divisions (e.g. node and edge sparsity in the network) and unified solutions with both new operations and problem transformations are encouraged to support a variety of sparse scenarios. These related studies may eventually settle more sparse challenging tasks in fields including recommendation systems, fault detections and social community studies.

Limitations of RiskSeq. In our accident prediction task, the classification error decreases when regression error increases due to the non-accident subregions are dominated

6. https://www.wunderground.com/history/daily/us/ny/new-york-city/KLGA/date/2017-4-22



(a) Three selected intervals on 22th, April, Sat, Cloudy to rainy

(b) Sequential results on 3rd, April, Mon, Rainy

Fig. 9: RiskSeq Visualization

in the city. Therefore, we should conduct further studies for a more concise, cost-sensitive model by equilibrating the tradeoffs between the pair of classification accuracy and regression capacity, as well as the pair of model complexity and interpretability. Another limitation is that RiskSeq still cannot fully avoid the over-predicting, and may reach its accuracy bottleneck. The accuracy bottleneck mostly pins on lacking the ability to sense the risks of individual vehicles. One possible solution is to collect individual statuses and improve local risk awareness with edge computing devices.

6 Related Work

In this section, we review related studies in three aspects, namely traffic accident prediction, classical time-series prediction and spatiotemporal traffic prediction.

Traffic accident prediction. Accident forecasting can be roughly folded into long-term forecasting [8], [12], [13], [23] and short-term forecasting [14], [15], [16], [17], [18], [24], [25], [26], [47]. Specifically, long-term forecasting methods model traffic-related data to predict the daily risks [2], [12], [13], [23]. For example, Chang et, al studied the highway accident frequency by a tree-based model on day levels [23]. Recently, Yuan et, al proposed a daily risk forecasting framework by employing a series of ConvLSTM sub-learners [13] and Huang et, al investigated to combine abnormal events and accidents to jointly predict future accidents in consecutive days with dynamic fusion network [12]. There has also been a citywide abnormal event forecasting framework proposed in [8], which shares similarity as accident predictions. Even so, all these daily forecasting models cannot support real-time traffic services and fail to incorporate unique characteristics between urban data and accident occurrences. Therefore, many efforts on short-term accident forecasting have been achieved [14], [15], [16], [17], [18], [24], [25], [26], [47]. Specifically, Lin et, al. formulated the task into a binary classification with frequent-pattern trees and random forest learning [24]. Some works quantified the spread of accident risks by employing Network Kernel Density Estimation and clustering methods [14], [15]. With the prosperity of deep learning, deep encoder-decoder mechanisms were introduced to satisfy the citywide risk predictions through stacking fully-connected layers and convolution blocks [17], [18], [25]. To deal with risk sequences and capture short-term temporal dependencies, some researchers tried to modify sequential learning schemes for accident predictions [16], [26], [47]. Unfortunately, above-mentioned approaches either model both spatial and temporal dependencies with traditional learning methods, or apply existing deep learning frameworks, hence all of them fail to identify distinctive observations in accident occurrences.

Classical time-series prediction. Forecasting accident risks can be viewed as time-series predictions. Off-the-shelf time-series predictions, like PSO-based methods [3], [4], [5], and ARIMA [6] can well capture temporal correlations and high-efficiently, but they mostly fail to address highly dynamic and complex road network traffic status due to their inherent linear or one-dimension fusions.

Spatiotemporal traffic prediction. Since traffic forecasting is well recognized to solve with spatiotemporal modelling, emerging researches proposed deep learning-based methods to jointly address spatial and temporal dependencies [7], [9], [10], [46]. These studies devised a series of methods such as diffusion convolution blocks and GCN-based sequence learning to foresee fundamental traffic elements and taxi demands in upcoming steps [10], [42], [46]. More recently, [9] proposed a meta-learning-based spatiotemporal forecasting method to increase the stability of transfer by learning common knowledge from multiple cities and [7] investigated a gate mechanism to model volume and flow interactions, which advance spatiotemporal forecasting community. However, the sporadic accident series with non-sufficient spatial sensing data are different from intensive and continuous sequences that can be trained without zero-inflated issue.

In summary, even though traditional optimization methods like PSO were efficient and many advanced

spatiotemporal deep learning methods like ConvLSTM and STDN have achieved promising results, none of above works raised the issue of the multi-step short-term accident forecasting, which is more challenging due to the sporadic spatial distribution and complex temporal tendency of traffic events. Therefore, these previous techniques were limited in addressing multi-granularity spatiotemporal accident forecasting.

7 Conclusion

In this paper, we propose a novel unified framework, RiskSeq, where sparse traffic accidents are learned with multiple spatiotemporal granularities, benefiting diversified requirements of travelers and traffic administrations. First, we summarize two kinds of sparsity challenges and correspondingly address these zero-inflated and sparse sensing issues. Inspired by the fresh observations in traffic accidents, we design a DT-GCN to enhance time-sensitive graph representations of risks by capturing short-term changes of urban traffics. To perform a multi-scale and multi-step prediction, coarse-grained and fine-grained risk distributions are learned simultaneously. With CG-LSTM, we can dynamically learn the region-context interactions and further alleviate the error accumulations. Experimental results on two real-world datasets prove the superiority of the integrated structure of DT-GCN and CG-LSTM in RiskSeq.

Future directions of the task-specific promotion is to leverage both global and local traffic information to maximumly reduce the individual random factors. The task-independent improvement comes down to further handle spatiotemporal sparsity issues for more general predictions.

8 Acknowledgements

This paper is partially supported by the Anhui Science Foundation for Distinguished Young Scholars (No.1908085J24), NSFC (No.61672487, No.61772492), Jiangsu Natural Science Foundation (No.BK20171240, BK20191193) and CAS Pioneer Hundred Talents Program.

References

[1] C. Pal, S. Hirayama, S. Narahari, M. Jeyabharath, G. Prakash, and V. Kulothungan, "An insight of world health organization (who) accident database by cluster analysis with self-organizing map (som)," *Traffic injury prevention*, vol. 19, no. sup1, pp. S15–S20, 2018.

[2] Tennessee model, "Tennessee traffic highway patrol," https://www.tn.gov/safety/news/2017/10/10/, 2017.

[3] L. Y. Hsu, S. J. Horng, T. W. Kao, Y. H. Chen, R. S. Run, R. J. Chen, J. L. Lai, and I. H. Kuo, "Temperature prediction and taifex forecasting based on fuzzy relationships and mtpso techniques," *Expert Systems with Applications*, vol. 37, no. 4, pp. 2756–2770, 2010.

[4] Y.-L. Huang, S.-J. Horng, T.-W. Kao, R.-S. Run, J.-L. Lai, R.-J. Chen, I.-H. Kuo, and M. K. Khan, "An improved forecasting model based on the weighted fuzzy relationship matrix combined with a pso adaptation for enrollments," *International Journal of Innovative Computing, Information and Control*, vol. 7, no. 7A, pp. 4027–4046, 2011.

[5] L. Y. Hsu, S. J. Horng, P. Fan, M. K. Khan, U. R. Wang, R. S. Run, J. L. Lai, and R. J. Chen, "Mtpso algorithm for solving planar graph coloring problem," *Expert Systems with Applications An International Journal*, vol. 38, no. 5, pp. 5525–5531, 2011.

[6] G. Zhang, "Time series forecasting using a hybrid arima and neural network model," *Neurocomputing*, vol. 50, pp. 159 – 175, 2003.

[7] H. Yao, X. Tang, H. Wei, G. Zheng, and Z. Li, "Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 5668–5675.

[8] C. Huang, C. Zhang, J. Zhao, X. Wu, D. Yin, and N. Chawla, "Mist: A multiview and multimodal spatial-temporal learning framework for citywide abnormal event forecasting," in *The World Wide Web Conference*, ser. WWW '19. New York, NY, USA: ACM, 2019, p. 717–728.

[9] H. Yao, Y. Liu, Y. Wei, X. Tang, and Z. Li, "Learning from multiple cities: A meta-learning approach for spatial-temporal prediction," in *The World Wide Web Conference*, 2019, pp. 2181–2191.

[10] L. Bai, L. Yao, S. Kanhere, X. Wang, Q. Sheng *et al.*, "Stg2seq: Spatial-temporal graph to sequence model for multi-step passenger demand forecasting," *arXiv preprint arXiv:1905.10069*, 2019.

[11] C. Song, Y. Lin, S. Guo, and H. Wan, "Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting," in *AAAI 2020 : The Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[12] C. Huang, C. Zhang, P. Dai, and L. Bo, "Deep dynamic fusion network for traffic accident forecasting," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, ser. CIKM '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 2673–2681.

[13] Z. Yuan, X. Zhou, and T. Yang, "Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 984–992.

[14] Anderson and K. Tessa, "Kernel density estimation and k-means clustering to profile road accident hotspots," *Accident Analysis & Prevention*, vol. 41, no. 3, pp. 359–364, 2009.

[15] Z. Xie and J. Yan, "Detecting traffic accident clusters with network kernel density estimation and local spatial statistics: an integrated approach," *Journal of transport geography*, vol. 31, pp. 64–71, 2013.

[16] H. Ren, Y. Song, J. Wang, Y. Hu, and J. Lei, "A deep learning approach to the citywide traffic accident risk prediction," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 3346–3351.

[17] C. Chen, X. Fan, C. Zheng, L. Xiao, M. Cheng, and C. Wang, "Sdcae: Stack denoising convolutional autoencoder model for accident risk prediction via traffic big data," in *2018 Sixth International Conference on Advanced Cloud and Big Data (CBD)*. IEEE, 2018, pp. 328–333.

[18] L. Zhu, T. Li, and S. Du, "Ta-stan: A deep spatial-temporal attention learning framework for regional traffic accident risk prediction," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[19] Z. Zhou, "Attention based stack resnet for citywide traffic accident prediction," in *2019 20th IEEE International Conference on Mobile Data Management (MDM)*. IEEE, 2019, pp. 369–370.

[20] E. O. Tufuor and L. R. Rilett, "Validation of the highway capacity manual urban street travel time reliability methodology using empirical data," *Transportation research record*, vol. 2673, no. 4, pp. 415–426, 2019.

[21] X. Zhou, Y. Shen, Y. Zhu, and L. Huang, "Predicting multi-step citywide passenger demands using attention-based neural networks," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2018, pp. 736–744.

[22] L. Bai, L. Yao, S. S. Kanhere, X. Wang, W. Liu, and Z. Yang, "Spatio-temporal graph convolutional and recurrent networks for citywide passenger demand prediction," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 2293–2296.

[23] L. Chang and W. Chen, "Data mining of tree-based models to analyze freeway accident frequency," *Journal of safety research*, vol. 36, no. 4, pp. 365–375, 2005.

[24] L. Lin, Q. Wang, and A. 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*, vol. 55, pp. 444–459, 2015.

[25] Q. Chen, X. Song, H. Yamada, and R. Shibasaki, "Learning deep representation from big and heterogeneous data for traffic accident inference," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[26] J. Bao, P. Liu, and S. V. Ukkusuri, "A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-

source data," *Accident Analysis & Prevention*, vol. 122, pp. 239–254, 2019.
[27] Z. Zhou, Y. Wang, X. Xie, L. Chen, and H. Liu, "Riskoracle: A minute-level citywide traffic accident forecasting framework," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34. AAAI, 2020, pp. 1258–1265.
[28] B. Wang, X. Luo, F. Zhang, B. Yuan, A. L. Bertozzi, and P. J. Brantingham, "Graph-based deep modeling and real time forecasting of sparse spatio-temporal data," in *MiLeTS 2018, London, United Kingdom*, 2018.
[29] Y. Wang, Y. Xiao, X. Xie, R. Chen, and H. Liu, "Real-time traffic pattern analysis and inference with sparse video surveillance information." in *IJCAI*, 2018, pp. 3571–3577.
[30] X. Geng, Y. Li, L. Wang, L. Zhang, Q. Yang, J. Ye, and Y. Liu, "Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting," in *2019 AAAI Conference on Artificial Intelligence (AAAI'19)*, 2019.
[31] J. Lian, X. Zhou, F. Zhang, Z. Chen, X. Xie, and G. Sun, "xdeepfm: Combining explicit and implicit feature interactions for recommender systems," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 1754–1763.
[32] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information theory*, vol. 37, no. 1, pp. 145–151, 1991.
[33] L. Chen, X. Fan, L. Wang, D. Zhang, Z. Yu, J. Li, T.-M.-T. Nguyen, G. Pan, and C. Wang, "Radar: Road obstacle identification for disaster response leveraging cross-domain urban data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 4, p. 130, 2018.
[34] Z. Liu, Z. Li, K. Wu, and M. Li, "Urban traffic prediction from mobility data using deep learning," *IEEE Network*, vol. 32, no. 4, pp. 40–46, 2018.
[35] S. M. Rifaat, R. Tay, and A. De Barros, "Effect of street pattern on the severity of crashes involving vulnerable road users," *Accident Analysis & Prevention*, vol. 43, no. 1, pp. 276–283, 2011.
[36] Z. Pan, Y. Liang, W. Wang, Y. Yu, Y. Zheng, and J. Zhang, "Urban traffic prediction from spatio-temporal data using deep meta learning," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 1720–1730.
[37] Y. Wang, Z. Zhou, K. Liu, X. Xie, and W. Li, "Large-scale intelligent taxicab scheduling: A distributed and future-aware approach," *IEEE Transactions on Vehicular Technology*, vol. 69, 2020.
[38] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks."
[39] Y. Zheng, H. Zhang, and Y. Yu, "Detecting collective anomalies from multiple spatio-temporal datasets across different domains," in *Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems*. ACM, 2015, p. 2.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2016.
[41] S. Abu-El-Haija, B. Perozzi, A. Kapoor, N. Alipourfard, K. Lerman, H. Harutyunyan, G. Ver Steeg, and A. Galstyan, "Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing," in *International Conference on Machine Learning*, 2019, pp. 21–29.
[42] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," in *International Conference on Learning Representations*, 2018.
[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference for Learning Representations*, 2015.
[44] D. Liao, W. Liu, Y. Zhong, J. Li, and G. Wang, "Predicting activity and location with multi-task context aware recurrent neural network." in *IJCAI*, 2018, pp. 3435–3441.
[45] N. Zhao and Z. Li, "Optimize traffic police arrangement in easy congested area based on improved particle swarm optimization," *Procedia-Social and Behavioral Sciences*, vol. 138, pp. 408–417, 2014.
[46] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. AAAI Press, 2018, pp. 3634–3640.
[47] S. Moosavi, M. H. Samavatian, S. Parthasarathy, R. Teodorescu, and R. Ramnath, "Accident risk prediction based on heterogeneous sparse data: New dataset and insights," in *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2019, pp. 33–42.

Zhengyang Zhou is now the master degree candidate in the School of Computer Science and Technology, University of Science and Technology of China. His research interests include machine learning, spatiotemporal data mining as well as artificial intelligence in traffic applications. He is a student member of AAAI and IEEE.

Yang Wang is now an associate professor at USTC. He got his Ph.D. degree at University of Science and Technology of China in 2007, under supervision of Professor Liusheng Huang. He also worked as a postdoc at USTC with Professor Liusheng Huang. His research interest mainly includes wireless (sensor) networks, distributed systems, data mining, and machine learning.

Xike Xie is currently a research professor in the School of Computer Science and Technology, University of Science and Technology of China. His research interests include data uncertainty, spatiotemporal databases, and mobile computing. He is a member of ACM and IEEE.

Lianliang Chen is now the master degree candidate in the School of Computer Science and Technology, University of Science and Technology of China. His research interests include machine learning, data mining as well as artificial intelligence in air quality.

Chaochao Zhu is now the master degree candidate in the School of Software Engineering, University of Science and Technology of China. His research interests include machine learning, data mining and artificial intelligence in traffic applications. He has achieved an excellent place in the KDD Cup 2019 competition.