A2DJP: A Two Graph-based Component Fused Learning Framework for Urban Anomaly Distribution and Duration Joint-Prediction

Kun Wang, Zhengyang Zhou, *Student Member, IEEE,* Xu Wang, Pengkun Wang, Qi Fang, and Yang Wang✉, *Member, IEEE*

Abstract—Modern intelligent transportation system (ITS) has greatly benefitted people's daily life. However, the chanciness and suddenness of urban anomalies may greatly restrict the trouble-free operations of ITS. To be aware of future urban anomalies and their possible influences, great efforts have been achieved on these two aspects, but comprehensive predictions of urban anomalies including the predictions of distributions and durations, are still beingless. And the spatiotemporal cascade self/mutual exciting influences among anomalies have never been considered in previous studies. In this paper, we propose a novel Anomaly Distribution and Duration Joint-Prediction (A2DJP) algorithm to simultaneously filtrate urban subregions and estimate the duration of corresponding potential anomalies in the future. To capture the spatiotemporal correlations between urban traffics and anomalies, we use a modified Graph Convolution Network and Long Short-Term Memory integrated network. To learn the cascade correlations among anomalies themselves, we devise a novel Spatiotemporal neural Hawkes Process model, which contains a Hawkes Process (HP) based GCN and HP-based LSTM to extract the anomaly-wise spatiotemporal cascading correlations. By fusing the spatiotemporal correlations between traffics and anomalies, we then simultaneously predict the distributions and durations of future anomalies. Extensive experiments on real-world datasets demonstrate that our proposed method significantly outperforms state-of-the-art solutions.

Index Terms—Hawkes Process, spatiotemporal cascading correlations, anomaly prediction, anomaly duration prediction.

1 Introduction

In recent years, to handle the increasingly serious traffic congestions and facilitate daily urban travels, Intelligent Transportation System (ITS) [1] was developed to provide high-quality intelligent transportation services based on the massive collection of traffic-related data and the effective management of urban traffics. However, urban anomalies, which include different kinds of abnormal urban events such as unusual congestions, accidental traffic accidents, and sporadic road obstructions, may bring great obstacles to the smooth operations of existing ITSs [2].

Simultaneously being aware of both distributions and durations of future urban anomalies [1] is of great significance to achieve intelligent transportations. Specifically, knowing potential urban anomaly distributions can reduce the likelihood of anomalies and smooth urban road networks by scheduling some corresponding resources in advance, and estimating feasible durations of future anomalies can directly rationalize the planning of urban trips and reduce the possibilities of secondarily derived anomalies.

Intuitively, the distributions and the durations of urban anomalies are mutually correlated with each other. For instance, only the places where accidents occur have the corresponding durations, and a place with a higher anomaly risk usually has a relatively higher anomaly severity and a longer average anomaly duration. Considering the strong correlations between the distributions and durations of urban anomalies and the great significance of simultaneously understanding both of them, the predictions of these two tasks should be viewed as a joint-prediction mission to force them

Fig. 1: An example of cascading correlations from anomaly to anomaly: at 18:30, due to heavy urban traffic during rush hours, an accident occurs at an intersection and cannot be resolved in time. Shortly, this accident consequently causes serious congestion and such congestion spread around along every out-going direction of the intersection in 5 to 10 minutes. With the aggravation of congestion, new anomaly is caused at 18:45.

- *K Wang, Z.Y. Zhou, X. Wang, P.K. Wang and Y. Wang are all with University of Science and Technology of China, Hefei, Anhui, P.R.China.*
- *Qi Fang is with University of Harbin Engineering, Harbin, Heilongjiang, P.R.China.*
- ✉ *Yang Wang is the corresponding author, E-mail: angyan@ustc.edu.cn.*

Manuscript received April 29, 2021.

1. The distributions and durations of future urban anomalies respectively indicate the specific numbers of anomalies and the average durations of anomalies within each specific sub-region of an urban area in the future.

to promote each other from the perspective of learning.

Great efforts have been achieved in understanding urban anomalies, and existing works can be divided into two categories: i) distribution prediction of urban anomalies [3], [4], [5], [6], [7], [8], [9], [10], [11], and ii) duration estimation of urban anomalies [12], [13], [14], [15], [16]. Regarding the first category of anomaly distributions, traditional deep learning methods mostly use Recurrent Neural Network based (RNN-based) or Convolutional Neural Network based (CNN-based) frameworks to capture the correlations between features and anomaly distributions. However, these traditional methods cannot effectively capture any dynamic or non-Euclidean correlation due to the time-invariant and localized characteristics of their embedded aggregators. Most recently, anomaly distribution methods are developed by time-varying GNNs, which are capable of capturing the dynamic non-Euclidean correlations among each region [10], [17], which are naturally suitable for graph-structured data, to fully extract spatiotemporal patterns between traffic features and anomalies within urban road network. However, merely predicting future accidents without giving the corresponding duration cannot accurately guide daily travel and traffic management. Obviously, the occurrence of an independent anomaly in a random road segment usually causes serious traffic congestions within surrounding areas. This accident may create a complex incentive relationship for the distribution and duration of surrounding traffic accidents. Figure 1 shows a real case which demonstrates cascading correlations from anomaly to anomaly. Regarding the second category, existing results mostly focus on the single task of predicting the durations of future anomalies by taking advantage of different kinds of regression analysis or temporal learning methods, and obviously, these duration prediction approaches are incapable of achieving the joint-prediction of both distributions and durations of future anomalies. In summary, to the best of our knowledge, none of existing works on the field of future anomaly prediction have considered the joint-prediction of distributions and durations of future anomalies, and the direct spatiotemporal cascading correlations [18], [19]. among anomalies have never been well captured in previous anomaly related predictions.

To address the above-mentioned issues, in this paper, we propose a novel Anomaly Distribution and Duration Joint-Prediction (A2DJP) algorithm by developing two graph-based pipelines to simultaneously filtrate urban subregions with high anomalous degrees in the future and estimate the duration of future possible anomalies. Specifically, for the GCN-LSTM pipeline, we employ a Graph Convolution Network (GCN) and Long Short-Term Memory (LSTM) network integrated network to embed urban traffic volumes in both spatial and temporal perspectives. Worth mentioning that there exist strong similarities between occurrence probabilities of anomalies and traffic flow density, so the fixed adjacency matrices of GCN are modified by traffic volume similarity matrices for capturing dynamic correlations between similar road segments. For the HP-GCN-LSTM pipeline, we propose a novel SpatioTemporal neural Hawkes Process (ST-HP) model, which borrows the core idea of accumulated and decayed influences from HP, to learn the direct cascading correlations among anomalies from both spatial and temporal perspectives. Finally, by fusing the spatiotemporal correlations among urban traffics and anomalies as well as the spatiotemporal cascading correlations among anomalies, we then simultaneously predict the distributions and durations of future anomalies. These two components are trained jointly in a multi-task learning manner and the design of each block has its scene and special function. The main contributions of this paper are as follows:

- We propose a discrete-continuous task to respectively achieve the predictions of discrete anomalies and continuous corresponding durations. To our best knowledge, this is the first work targeting the joint-prediction issue of future urban anomaly distributions and durations. We use a novel two pipelines fusion network to decouple the entangled influences of urban traffic volumes and direct cascading correlations to future anomalies respectively in both spatial and temporal perspectives. In order to capture the correlation between the two sub-tasks, we use a negative cosine loss function to constrain the similarity of the two outputs and simultaneously filtrate subregions with high anomalous degrees and predict the duration of these anomalies.
- We design a novel ST-HP, which contains two novel deep learning variants, HP-GCN and HP-LSTM, to respectively capture the spatial and temporal cascading non-Euclidean interactions among anomalies. By utilizing and extended the core idea of Hawkes Process to concern the inherent self/mutual-exciting characteristics of anomalies, ST-HP can effectively capture the direct cascading correlations among anomalies.
- We evaluate our proposed framework via two real-world datasets, New York open data [20] and US_Datasets [21], and extensive experiments demonstrate our propose method outperforms other alternative approaches in terms of the accuracies of the distribution and duration joint-prediction. Specifically, compared with the state-of-the-art single-task anomaly distribution prediction and duration estimation solutions respectively, our method gains at least a 2.2% increase in both the predictions of anomaly distributions and durations with different datasets in terms of Acc@20.

2 Related work

Great efforts have been studied in addressing the issue of future anomaly prediction, and most existing works [3], [4], [5], [6], [7], [8], [9], [10], [11] about this issue focus on predict distributions of future urban anomalies, and the rest few works [12], [13], [14], [15], [16] concern about the duration estimation of urban anomalies.

The issue of predicting distributions of future anomalies can also be seemed as predicting future anomalies. Regarding this issue, [3] first proposes a classification and regression tree and a negative binomial regression model to construct the correlations between anomalies and geometric features of road segments for predicting future anomalies, [5] proposes a support vector machine with Gaussian kernel to predict the occurrence probabilities of anomalies, and [4] devises a nonnegative matrix factorization based model to predict risk levels of anomalies. However, these traditional machine learning based methods are limited in approximating complex spatiotemporal correlations from multi-source data. To this end, some deep learning based methods including RNN based model [7] and CNN based model [6] are proposed to capture complex spatiotemporal dependencies. Regarding these methods, they usually stack multi-layer aggregators to capture dependencies between long range neighbors, and this leads to

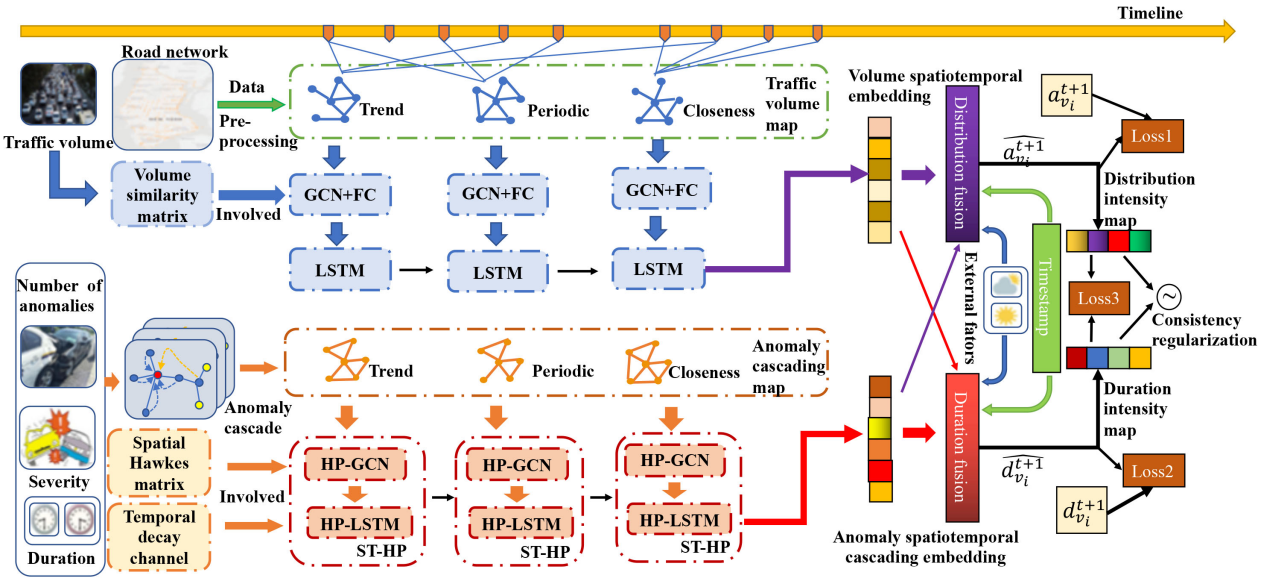

Fig. 2: Solution overview.

the tremendous computational complexities of proposed methods. Further, traditional grid-based deep learning models cannot extract any dynamic spatiotemporal correlation due to their inherent invariant characteristics. To address these challenges, in recent years, GNN based networks such as GCN [22] and Graph Attention Network (GAT) [23] are used to predict anomalies in graph structured urban road networks. In particular, [8] proposes a temporal GCN, which integrates gate recurrent unit with GCN, to effectively detect abnormal traffic volumes within urban road network by capturing both spatial and temporal correlations between urban traffic features and volumes. [9] devises a novel relational GCN to predict failure paths within road network, [10] uses a novel differential time-varying GNN to predict traffic accidents by capturing the differences of vehicle numbers and traffic volumes in time series. Based on GAT, [11] proposes a novel graph multi-attention network, which contains multiple spatio-temporal attention blocks, to detect traffic volume and speed anomalies within road network. In summary, some events are excited or inhibited by patterns in the sequence of previous events, none of them has considered the direct cascading spatiotemporal interactions among anomalies themselves. And nevertheless, existing methods on predicting distributions of future anomalies cannot be directly used to address the issue of anomaly duration prediction.

Regarding predicting anomaly durations, some machine learning based methods are proposed firstly. [12] uses a hybrid tree-based quantile regression model to quantify the influences of anomaly categories and traffic features on anomaly durations, [13] proposes a K-nearest neighbor based model to address the issues of sample-disequilibrium and cost-sensitive in predict durations of anomalies, and [14] uses Xgboost to comprehensively estimate the impact of each individual traffic feature on durations of future anomalies. However, these machine learning based methods can only use features of anomalies to predict anomaly durations, and they have never taken the spatiotemporal information of anomalies into account. To this end, recently, deep learning models are then considered to be used to address this issue. To integrate spatiotemporal information of anomalies in predicting anomaly durations, [15] uses restricted Boltzmann machines and [16] devises a novel scaled multinomial logit model. Without exception, all these methods can only be used to predict durations of future anomalies, and none of them is capable of predicting anomaly distributions. And likewise, none of these methods has considered the direct cascading correlations among anomalies themselves.

In summary, none of the existing works has raised the joint-prediction issue of both anomaly distributions and durations, and the direct cascading correlations among anomalies have never been distinctively decoupled and considered in both spatial and temporal perspectives either.

3 Problem Definition

In this section, we first introduce some preliminaries and basic settings about this paper, and then formally define the problem that we discussed in this paper. For convenient modeling in subsequent steps, we first divide the total urban area into an $I \times J$ grid, and each individual grid point is a square with the length of l. Worthing note that the setting of l should equilibrate the trade-off between the practicability and spatial granularity. In our implementation, we divide the whole urban areas of New York City and Chicago into small squares with the length of 1.5 km following the requirements of practicability, and the urban areas of these two cities are partitioned into 30×22 and 28×19 grids respectively. Given the grid division, the whole urban area can be denoted as a subregion set $\mathcal{V} = \{v_1, v_2, \cdots, v_N\}$ where $N = I \times J$. For one specific grid point v_i, we then define its time-varying input features, i.e.,

Definition 1 (Time-varying input features of subregions). *The setting of the length of an interval should trade-off between the prediction performances and temporal granularity. In our implementation, we slice the temporal information into intervals of 30 minutes following the common settings. Notice that such a setting may be related to the results of prediction performance. Regarding subregion v_i and time interval t, the corresponding time-varying feature set can be denoted by*

$$\mathcal{F}_{v_i}^t = \{f_{v_i}^t, a_{v_i}^t, s_{v_i}^t, d_{v_i}^t\} \quad (1)$$

TABLE 1: Notation of features

Notation	Definition
$f_{v_i}^t$	Traffic volume of v_i during t
$a_{v_i}^t$	Total number of anomalies in v_i during t
$s_{v_i}^t$	Severity of anomalies in v_i during t
$d_{v_i}^t$	Duration of anomalies in v_i during t

where notice here the traffic volume of a sub-region during a specific interval corresponds to the average traffic volume of all road segments within this grid area during that interval. The anomaly severity of $s_{v_i}^t$ can be calculated by

$$s_{v_i}^t = \sum_j s_{v_i}^t(j) \tag{2}$$

where $s_{v_i}^t(j)$ is the severity of the jth anomaly in subregion v_i during interval t. And the duration of anomalies in v_i can be computed by

$$d_{v_i}^t = \sum_j d_{v_i}^t(j)/a_{v_i}^t \tag{3}$$

where $d_{v_i}^t(j)$ is the duration of the j-th anomaly in subregion v_i during interval t. $a_{v_i}^t$ is the total number of the anomalies in subregion v_i during interval t.

Definition 2 (Anomaly distribution and duration joint prediction problem). *Given $\mathcal{F}_{v_i}^t$ for all subregions $v_i \in \mathcal{V}$ and historical intervals $\{t-m+1, \cdots, t-1, t\}$, combining with some external meteorological factors including dew, body temperature, humidity, pressure, visibility, wind speed, gust speed, and weather condition, the main target of this problem is to filtrate the subregions with high incidences of accidents during the next interval $t+1$, i.e., $\eth_{t+1} = \{v_i \,|a_{v_i}^{t+1} > \beta, i \in [1, N]\}$ where β is a pre-defined threshold, and simultaneously predict the corresponding durations of anomalies in all regions within $\eth_{t+1}$. Here m is the length of historical temporal window for predicting future anomaly distribution and duration.*

4 Anomaly Distribution and Duration Joint-Prediction

In this section, we first introduce the overview of the proposed two graph-based component fused solution, and then describe the detailed implementation of each individual component.

4.1 Solution overview

In this paper, we propose a two-component-fused network to simultaneously filtrate urban subregions with high anomalous degrees in the future and estimate the duration of future possible anomalies. The overview of this two-component fused network is illustrated in Figure 2. As illustrated, in the first component, we employ a modified GCN and LSTM network-integrated component to embed urban traffic volumes in spatiotemporal perspective for learning the spatiotemporal correlations between urban traffics and anomalies, and in the second pipeline, we propose a novel ST-HP model to extract the direct cascading spatiotemporal correlations among anomalies. To realize the collaboration and mutual assistance of information, these two components are fused in a multi-task learning framework. Eventually, by introducing the bridge mechanism, we filter the learned node values to match the areas with both events and durations in one-to-one manners, and then suppress the duration of the accident-free area to zero. In the following parts, we then describe the detailed designs and implementations of the whole network.

4.2 Data pre-processing and feature map construction

In this subsection, we discuss the pre-processing issue of input features, and then discuss the construction of corresponding feature maps for the two different components.

As discussed, the input features of a specific sub-region v_i during an interval t include its traffic volume $f_{v_i}^t$, total anomaly number $a_{v_i}^t$, anomaly severity $s_{v_i}^t$ and average anomaly duration $d_{v_i}^t$. Our target is to predict the distributions of future anomalies and the corresponding durations. However, since the vast majority are zero-labeled data inputs, deep learning methods will definitely suffer from the notorious zero-inflated issue [24], [25]. For instance, if more than 90% of labels are zero, deep learning networks can achieve a nearly 90% accuracy even if the predicted results are all zero. This zero-inflated problem may shield the performance and ability differences between different networks. For anomaly prediction research line, the two anomaly related labels, $a_{v_i}^t$ and $d_{v_i}^t$, are inherently sparse due to the sparsity of urban anomalies. Figure 3 illustrates and verifies the sparse distributions of anomalies in both New York City and Chicago during an interval, i.e., 30 minutes. This kind of sparsity is quite usual for most spatiotemporal datasets and may inherently cause the zero-inflated issue during training. To this end, we should address the sparse issue of these two anomaly-related labels firstly, and we here design and employ a Global Prior Knowledge-based Data Enhancing (GPKDE) strategy. Through this logarithm transformation, the discrete and sparse data based prediction becomes an explicit regression mission.

By taking the label of total anomaly number as an example, regarding a specific region v_i in region division $\mathcal{V}$, we first calculate the average anomaly risk of sub-region v_i for all possible time intervals t in the dataset by

$$\varepsilon_{v_i} = \frac{\sum_t a_{v_i}^t}{\sum_t \sum_{j=1}^{N} a_{v_j}^t} \tag{4}$$

notice here the total number of anomalies of a region during all possible intervals is normalized to a value within $[0, 1]$, and we subsequently calculate the statistical anomaly occurrence number of subregion v_i by,

$$\pi_{v_i} = b_1 \log_2(\varepsilon_{v_i} + \Delta) + b_2 \tag{5}$$

With the logarithm transformation, we can easily transform the average anomaly risk of sub-region v_i into a negative value, and here b_1 and b_2 are the coefficients to maintain that the range of the absolute value of π_{v_i} is the same to the range of anomaly numbers, hence preserving the ranks of actual anomaly numbers among sub-regions. Notice here Δ is an extremely small value to make sure that the calculation of the logarithm transformation can be trouble-free in case that $\varepsilon_{v_i} = 0$. For all sub-regions, i.e., $i \in [1, N]$, we use this statistical anomaly occurrence number π_{v_i} to replace $a_{v_i}^t$ in case that $a_{v_i}^t = 0$, and a subregion with lower statistical anomaly occurrence number has lower anomaly risk.

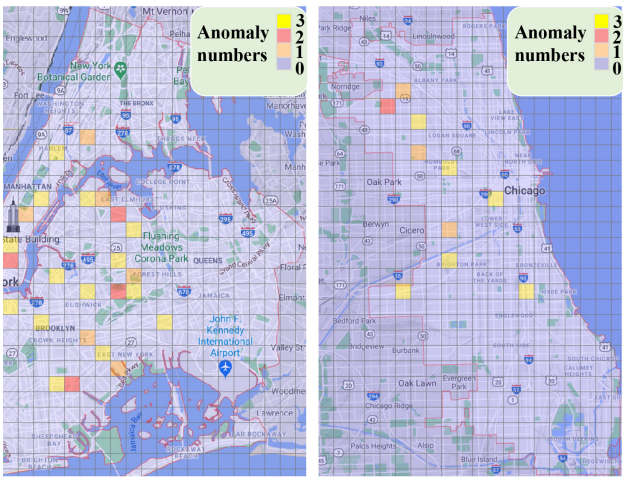

Fig. 3: Sparse distributions of anomalies in New York City and Chicago during an interval (30 minutes): the color filled in a grid demonstrates the number of anomalies within this subregions during 30 minutes, and the color of light blue, which can be found in most subregions, means there doesn't exist any anomaly during the 30 minutes.

In a similar way, regarding the label of anomaly duration, we first calculate the the average anomaly duration of sub-region v_i for all possible time interval t in dataset by,

$$\phi_{v_i} = \frac{\sum\limits_{t} d^t_{v_i}}{\sum\limits_{t}\sum\limits_{j=1}^{N} d^t_{v_j}} \tag{6}$$

and we subsequently calculate the statistical anomaly duration of subregion v_i by,

$$\theta_{v_i} = c_1 \log_2(\phi_{v_i} + \Delta) + c_2 \tag{7}$$

where c_1 and c_2 are also the coefficients to maintain that the range of the absolute value of θ_{v_i} is the same to the range of anomaly durations.

4.3 Modified GCN and LSTM integrated network for spatiotemporal learning

In this subsection, we use a modified GCN and LSTM integrated network to respectively capture the embedded representations of both spatial and temporal dependencies among traffic volumes. For extracting these spatiotemporal dependencies, we first construct the corresponding traffic volume map based on urban traffic volume dataset.

Construction of traffic volume map: Since urban traffic volumes follow obvious daily and weekly periodicities [26], we should construct the corresponding traffic volume map from three different granularities: closeness, periodicity, and trend. Given the traffic volumes of the entire subregion set $\mathcal{V}$, the traffic volume map of interval t can be denoted as $\mho^t = \{f^t_{v_1}, \cdots f^t_{v_N}\}$. Regarding time interval t, the traffic volume map in close granularity includes the traffic volume maps during a number of previous intervals before t, i.e., $\{\mho^{t-m+1}, \cdots \mho^t\}$, the traffic volume map in periodic granularity includes the traffic volume maps of the same interval with t in a day during a number of previous days, i.e., $\left\{\mho^{t-(m-1)\times 48}, \cdots, \mho^{t-48}, \mho^t\right\}$, and the traffic volume map in trendy granularity includes the traffic volume maps of the same interval with t in the same day with the current day in a week during a number of previous weeks, i.e., $\left\{\mho^{t-(m-1)\times 48\times 7}, \cdots, \mho^{t-48\times 7}, \mho^t\right\}$. With these traffic volume maps in three different granularities, deep learning model can effectively exploit both long-term and short-term periodicities.

Spatial learning with modified GCN: In recent years, GCN and its variants [27], [28], which have shown strong abilities to capture non-Euclidean correlations within graph-structured data, are suitable for mining road network related data. We here employ a modified GCN, whose adjacency matrix is replaced by a traffic volume similarity matrix, to adaptively learn the complex spatial dependencies of road network and aggregate the information of similar sub-regions. Specifically, GCN model follows layer-wise propagation, i.e.,

$$H^{(l+1)} = \sigma(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^l W^l) \tag{8}$$

where $\widetilde{A}$ is the adjacency matrix of graph with added self-connections, $\widetilde{D}$ corresponds to the degree matrix for $\widetilde{A}$. Here σ denotes an activation function, and W^l is the learnable weight matrix in the l-th layer, H^l is the l-th hidden layer of GCN, and H^0 indicates the input feature matrix of GCN. Given the fact that there exist obvious similarities between the occurrence probabilities of anomalies within two urban sub-regions if their traffic volumes are also similar [10], [29], we here design a novel traffic volume similarity matrix to imply the potential anomaly pattern. Specifically, given two sub-region v_i and v_j, the traffic volume similarity between these two sub-regions with regard to interval t can be calculated with cosine similarity, i.e.,

$$\psi_t(\mathbb{F}^t_{v_i}, \mathbb{F}^t_{v_j}) = \cos(\mathbb{F}^t_{v_i}, \mathbb{F}^t_{v_j})^{\,2} \tag{9}$$

where $\mathbb{F}^t_{v_i} = \{f^{t-m+1}_{v_i}, \cdots, f^t_{v_i}\}$ indicates the traffic volumes of sub-region v_i during a temporal window with fixed length m. And the similarity matrix of interval t can be written as

$$\Psi_t = \begin{Bmatrix} \psi_t(\mathbb{F}^t_{v_1}, \mathbb{F}^t_{v_1}) & \cdots & \psi_t(\mathbb{F}^t_{v_1}, \mathbb{F}^t_{v_N}) \\ \vdots & \ddots & \vdots \\ \psi_t(\mathbb{F}^t_{v_N}, \mathbb{F}^t_{v_1}) & \cdots & \psi_t(\mathbb{F}^t_{v_N}, \mathbb{F}^t_{v_N}) \end{Bmatrix} \tag{10}$$

With this similarity matrix, the GCN module can be modified as,

$$H^{(l+1)} = \sigma(\widetilde{D}^{-\frac{1}{2}}_{\Psi_t}\Psi_t\widetilde{D}^{-\frac{1}{2}}_{\Psi_t}H^l W^l) \tag{11}$$

where H^0 corresponds to the traffic volume maps in three different granularities, and $\widetilde{D}_{\Psi_t}$ corresponds to the degree matrix for Ψ_t. Worth noting that the computational complexity of training this modified GCN for all possible interval is $\mathrm{O}(N^2 * T)$ where T indicates the total number of possible intervals in the dataset. To equilibrate the contradiction between the calculation burden and performance of our algorithm, we set a traffic volume similarity threshold ξ, and those cosine similarity values within two subregions, which are less than the threshold ξ, are suppressed to 0.

Temporal learning with LSTM: We then feed the embedded representations of spatial dependencies learned by the modified GCN to the LSTM model to exploit the temporal dependencies among traffic volumes. LSTM can automatically control the weights of input information through the mechanism of "gate", and finally export the embedded representations of both spatial

2. $\cos(\overrightarrow{x_1}, \overrightarrow{x_2})$ indicates the cosine similarity between two vectors $\overrightarrow{x_1}$ and $\overrightarrow{x_2}$, and can be calculated by $\cos(\overrightarrow{x_1}, \overrightarrow{x_2}) = \frac{\overrightarrow{x_1}\cdot\overrightarrow{x_2}}{\|\overrightarrow{x_1}\|\times\|\overrightarrow{x_2}\|}$

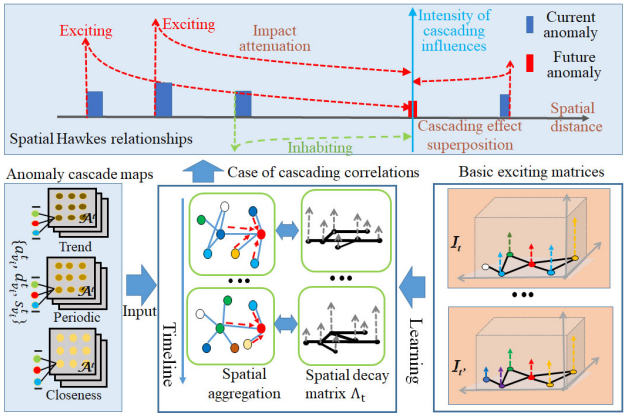

Fig. 4: HP-GCN for anomaly cascade spatial learning.

and temporal dependencies among traffic volumes. Due to the limitation of space, we here omit the detailed implementations of standard LSTM in this subsection.

4.4 ST-HP for anomaly cascade learning

HP is widely used to model the inherent self/mutual-exciting characteristics of social events [30] in previous studies, a traditional HP supposes that past event can temporarily raise the probability of future events, assuming that exciting characteristics is (1) positive, (2) additive over the past events, and (3) exponentially decaying with time. In recent years, some studies have tried to expand HP's functions and make it more in line with the patterns (excited or inhabited) of events in the real-world, and some even try to combine HP with neural network to learn the complex cascading correlations of events in time series [31], [32]. Considering the inherent spatiotemporal self/mutual-exciting characteristic of urban anomalies, to effectively capture the cascading correlations among anomalies, we here introduce spatiotemporal HP to modify traditional GCN and LSTM in both spatial and temporal perspectives. In this subsection, we here also first construct the corresponding anomaly cascading map based on urban anomaly dataset, then introduce the detailed implementations of the modified HP-GCN and HP-LSTM.

Construction of anomaly cascading map: Based on the features about anomalies including $a^t_{v_i}$, $s^t_{v_i}$, and $d^t_{v_i}$ in $\mathcal{F}^t_{v_i}$, we then construct the anomaly map of the entire urban area during interval t as $\mathcal{A}^t = \{\{a^t_{v_1}, s^t_{v_1}, d^t_{v_1}\}, \cdots, \{a^t_{v_N}, s^t_{v_N}, d^t_{v_N}\}\}$. Similar to the construction of traffic volume map, the corresponding anomaly cascading map should also be constructed in the granularities of closeness, periodicity, and trend. Regarding interval t, the anomaly cascading maps in close, periodic, and trendy granularities should be $\{\mathcal{A}^{t-m+1}, \cdots \mathcal{A}^t\}$, $\left\{\mathcal{A}^{t-(m-1)\times 48}, \cdots, \mathcal{A}^{t-48}, \mathcal{A}^t\right\}$, and $\left\{\mathcal{A}^{t-(m-1)\times 48\times 7}, \cdots, \mathcal{A}^{t-48\times 7}, \mathcal{A}^t\right\}$ respectively. Also, with these anomaly cascading maps in different granularities, deep learning model can effectively exploit both long-term and short-term periodicities within anomalies.

Spatial learning with HP-GCN: As discussed, an anomaly may propagate its influences to surrounding areas, hence the occurrence of new anomalies or the increasing of severities of surrounding anomalies. On the other hand, this kind of influence decrease with the increase of the distance within anomalies. To capture this kind of cascading spatial correlations among anomalies, we import HP into GCN. A detailed demonstration about how HP-GCN learns cascading correlations among anomalies in spatial perspective is given in Fig. 4. Specifically, given a subregion $v_i \in \mathcal{V}$, to model the influences of anomalies within this grid point on other surrounding subregions, we first define the basic exciting influence of v_i during interval t by

$$I^t_{v_i} = \begin{cases} \frac{s^t_{v_i}}{a^t_{v_i}} & a^t_{v_i} \neq 0 \\ 0 & a^t_{v_i} = 0 \end{cases} \tag{12}$$

Notice here we use the quotient of the severity $s^t_{v_i}$ over the total anomaly number $a^t_{v_i}$ as the basic exciting influence of v_i during t in case that there exist anomalies within v_i during t. Therefore, the basic exciting influence matrix of $\mathcal{V}$ during interval t can be formalized by

$$\mathcal{I}_t = \begin{bmatrix} I^t_{v_1} & \cdots & I^t_{v_1} \\ \vdots & \ddots & \vdots \\ I^t_{v_N} & \cdots & I^t_{v_N} \end{bmatrix} \tag{13}$$

Notice that in Fig. 4, the upturned arrows indicate the basic exciting influence of each individual subregion. On the other hand, to calculate the decayed influences between two specific subregions, we design a decaying matrix by

$$\mathcal{D} = \begin{bmatrix} e^{\frac{-\eta_{v_1 v_1}}{\eta_{max}}} & \cdots & e^{\frac{-\eta_{v_1 v_N}}{\eta_{max}}} \\ \vdots & \ddots & \vdots \\ e^{\frac{-\eta_{v_N v_1}}{\eta_{max}}} & \cdots & e^{\frac{-\eta_{v_N v_N}}{\eta_{max}}} \end{bmatrix} \tag{14}$$

where $\eta_{v_i v_j}$ indicates the Euclidean distance between the centers of v_i and v_j and η_{max} corresponds to the maximum Euclidean distance among the centers of all sub-regions. We here use the variable $e^{\frac{-\eta_{v_i v_j}}{\eta_{max}}}$ to evaluate the decaying factor of influences between these two specific subregions, and this factor increases in case that the distance between two subregions decreases, and the influence factor of a subregion on itself is 1 since $\eta_{v_i v_i} = 0$, and the influence factor equals to 0 while the distance between two subregions is ∞. So far, the decayed anomaly influences of all urban subregions pairs during interval t can be calculated by

$$\Lambda_t = \mathcal{I}_t \circ \mathcal{D} \tag{15}$$

where $\circ$ means Hadamard product. So far, the HP-GCN module can be formulated by,

$$H^{(l+1)} = \sigma(\widetilde{D}^{-\frac{1}{2}}_{\Lambda_t} \Lambda_t \widetilde{D}^{-\frac{1}{2}}_{\Lambda_t} H^l W^l) \tag{16}$$

where H^0 corresponds to the anomaly cascading maps in three different granularities, and $\widetilde{D}_{\Lambda_t}$ corresponds to the degree matrix for Λ_t. Regarding HP-GCN, the calculation of such decayed anomaly influence matrix is based on the basic principle of HP theory, and the calculated Λ_t is utilized as the adjacency matrix of HP-GCN, hence involving mutual spatial excitements among subregions in spatial learning. Moreover, as demonstrated in Equation (16), HP-GCN multiple Λ_t with the cascading feature matrix in each layer, and the result of such matrix multiplication at the first layer of our GCN model indicates the self/mutual influences for demonstrating

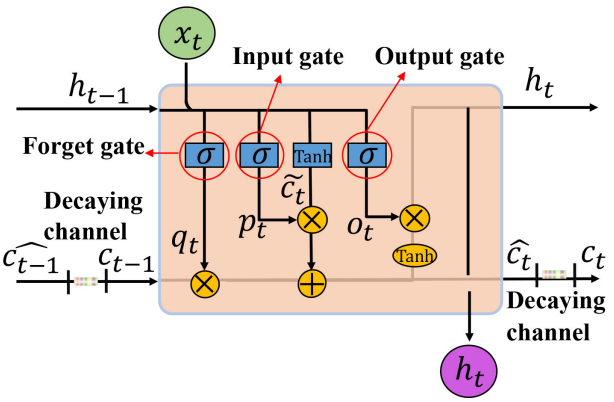

Fig. 5: Architecture of HP-LSTM.

the accumulated influences of all subregions on each subregion, i.e.,

$$\Lambda_t \cdot \begin{bmatrix} a^t_{v_1} \\ \vdots \\ a^t_{v_N} \end{bmatrix} = \begin{bmatrix} I^t_{v_1} \cdot \left(a^t_{v_1} e^{\frac{-\eta_{v_1 v_1}}{\eta_{max}}} + \cdots + a^t_{v_N} \cdot e^{\frac{-\eta_{v_1 v_N}}{\eta_{max}}} \right) \\ \vdots \\ I^t_{v_1} \cdot \left(a^t_{v_1} e^{\frac{-\eta_{v_1 v_1}}{\eta_{max}}} + \cdots + a^t_{v_N} \cdot e^{\frac{-\eta_{v_1 v_N}}{\eta_{max}}} \right) \end{bmatrix} \tag{17}$$

Here the i-th row of the result is the cumulative impact of anomaly numbers of all subregions on the number of anomalies in the i-th subregion. Likewise, the impact of all subregions on one specific subregion in terms of anomaly severity and duration can be calculated in the same way. With HP-GCN, the cascading correlations among anomalies can be extracted as shown in the upper half part of Fig 4, and the impact of cascading influences of current anomalies on a specific future anomaly are summed. Notice that the learned cascading influences from current anomalies to future anomalies can be bidirectional which means the influence of current anomalies on a future anomaly at a specific location can be either exciting or inhibiting, and we will further discuss the reason in the Section of "discussion".

Temporal learning with HP-LSTM: The output matrices of HP-GCN have already included the spatial representations of cascading correlations among anomalies, and we still have to capture the temporal cascading correlations among anomalies. To this end, we propose a novel HP-LSTM by integrating HP with LSTM. Figure 5 demonstrates the detailed implementation of HP-LSTM, and this HP-LSTM can be formalized by

$$x = \begin{cases} q_t = \sigma(W_q[h_{t-i}, x_t] + b_q) & (a) \\ p_t = \sigma(W_p[h_{t-i}, x_t] + b_q) & (b) \\ \widetilde{c}_t = \text{Tanh}(W_c[h_{t-1}, x_t] + b_c) & (c) \\ \widehat{c}_t = q_t \times c_{t-1} + p_t \times \widetilde{c}_t & (d) \\ 0_t = \sigma(W_o[h_{t-1}, x_t] + b_o) & (e) \\ c_t = W_{in} \cdot \widehat{c}_t \cdot \exp(\frac{-t_{inter}}{\rho_t}) & (f) \\ h_t = o_t \times \text{Tanh}(\widehat{c}_t) & (g) \end{cases} \tag{18}$$

where function $[,]$ means the concatenation of two vectors, $W_* = \{W_q, W_p, W_c, W_o,\}$ and $b_* = \{b_q, b_p, b_c, b_o,\}$ are learnable weights and offsets. Here "**Forget gate**" determines how much information in the last cell i.e., c_{t-1}, can be maintained into the cell of the current time, i.e., $\widehat{c}_t$, "**Input gate**" determines how much information of the current input x_t can be filter to the current cell $\widehat{c}_t$, and $\widehat{c}_t$, which indicates the combination of the filtered information of c_{t-1} and x_t, is outputted by "**Output gate**". Regarding HP-LSTM, different from traditional LSTM which directly fuse the outputs of forget gate and input gate as output, HP-LSTM regards the output $\widehat{c}_t$ of each cell as the basic exciting influence, and quantifies the temporal decaying influences of previous events on future events in Sub-Equation (f) of Equation (18) by referring to classic HP theory. Notice here W_{in} and ρ_t are both learnable parameters, and such modification on LSTM enables the adaptive learning of the temporal influence of previous events on future events in all three temporal granularities. Likewise, in HP-LSTM, the output of the forget gate corresponds to the self/mutual influence function which demonstrates the accumulated influences of anomalies during historical intervals on future anomalies.

4.5 Multi-task learning for 2D joint-prediction

The occurrence and severity of urban anomalies can be impacted by their occurring locations and time, external meteorological and urban traffic factors, as well as the cascading influences among anomalies. To fully fuse all these related information, we propose two novel fusion models: distribution fusion and duration fusion. Regarding distribution fusion, we fuse these information to predict the anomaly distribution of the next interval with a Fully Connected (FC) layer and calculate the regression loss of anomaly distribution prediction by

$$Loss_{dis} = \sum_t \sum_{i=1}^{N} \text{Dist}_{\text{dis}} \left(a^{t+1}_{v_i}, \widehat{a^{t+1}_{v_i}} \right) \tag{19}$$

Regarding duration fusion, we fuse all these information to predict the anomaly duration of the next interval with a FC layer, and the regression loss for this part can be written as,

$$Loss_{dur} = \sum_t \sum_{i=1}^{N} \text{Dist}_{\text{dur}} \left(d^{t+1}_{v_i}, \widehat{d^{t+1}_{v_i}} \right) \tag{20}$$

Here $\text{Dist}\,(y, \widehat{y})$ is a distance related function which can be used to quantify the difference between y and $\widehat{y}$. Notice this function can be Mean Absolute Error (MAE) or Mean Squared Error (MSE), and we then discussed the selection of this function in the section of "Experimental Studies". To better capture the correlations between these two different prediction missions, we add a negative cosine similarity loss to keep the two outputs from different parts as consistent as possible, i.e.,

$$Loss_{cos} = -\sum_t \frac{\widehat{\alpha^{t+1}} \cdot \widehat{\delta^{t+1}}}{||\widehat{\alpha^{t+1}}|| \cdot ||\widehat{\delta^{t+1}}||} \tag{21}$$

where $\widehat{\alpha^{t+1}} = \left\{ \widehat{a^{t+1}_{v_1}}, \cdots \widehat{a^{t+1}_{v_N}} \right\}$ and $\widehat{\delta^{t+1}} = \left\{ \widehat{d^{t+1}_{v_1}}, \cdots \widehat{d^{t+1}_{v_N}} \right\}$. So far, the total loss in training phase for 2D joint-prediction multi-task learning can be written as,

$$Loss(\Theta) = Loss_{dis} + \lambda_1 * Loss_{dur} + \lambda_2 * Loss_{cos} + \gamma \|\Theta\|^2 \tag{22}$$

where Θ includes all learnable parameters, $\|\cdot\|^2$ is L2-norm for preventing overfitting. We here adopt Adam optimizer to train our joint-prediction framework, design a novel consistency regularization architecture for mutual matching the predicted distributions and durations of all subregions within $\mathcal{V}$, and suppress the anomaly numbers and durations of those subregions both to 0 in case that their anomaly numbers are less than the threshold β meanwhile.

TABLE 2: Dataset statistics.

City	Dataset	Time Span	# of Regions	# of Records
New York City	Crashes/Accidents	01/01/2018-12/31/2018	30*22 (660)	204k
	Taxi Trip Data			6208k
	Meteorology			204k
	Anomaly duration			8.6k
Chicago	Crashes/Accidents	07/29/2018-12/31/2018	28*19 (532)	24k
	Taxi Trip Data			220k
	Meteorology			24k
	Anomaly duration			12.4k

TABLE 3: Parameter settings.

Symbol	Description	Value
$--$	Length of time intervals	30 min
m	Number of historical intervals	8
β	Threshold for filtrating anomalies	0.8
(Δ, b_1, b_2)	Overcome zero-inflated parameters 1	$(10^{-6}, 0.13, 0.66)$
(c_1, c_2)	Overcome zero-inflated parameters 2	$(0.12, 0.26)$
ξ	Threshold of volume similarity matrix	0.3
(λ_1, λ_2)	Weights of loss function	$(1.2, 0.8)$
$--$	Number of GCN blocks	4
$--$	Number of LSTMs	2

5 EXPERIMENTAL STUDIES

In this section, we evaluate the performances of the proposed model, A2DJP, on the urban anomaly datasets of New York city and Chicago. Meanwhile, we conduct extensive ablation experiments to verify the effectiveness of each component. Eventually, to further intuitively analyze our model, we introduce an interesting case study.

5.1 Data preparation

Our experiments are conducted based on the real-world anomaly distribution and duration datasets of two different cities, New York City and Chicago. Regarding these two datasets, urban anomalies contain unusual congestion, accidental traffic accidents, and sporadic road obstructions. Regarding New York city, the datasets are during January 1, 2018 and December 31, 2018. For Chicago, the datasets are within August 1, 2018 and December 31, 2018. For both two cities, we utilize the taxicab volumes within subregions as the indicator of traffic volumes of all subregions. The statistics of data are shown in Table 2.

5.2 Experimental settings

In our experiments, we select 70%, 20% and 10% of datasets for training, evaluation and validation. Meanwhile, to eliminate the dimensional influences between different indicators, we normalize all the features in the datasets. The New York City and Chicago are both divided into squared subregions with the length of $1.5km \times 1.5km$. During the training phase, we set the batch size at 64 and the learning rate at 0.001. Eventually, we implement our model, A2DJP, based on Pytorch[3]. All parameter settings related to our model are summarized in Table 3.

To evaluate the performance of our proposed model and all alternative solutions, we use the information of urban traffics and anomalies during m previous intervals to predict the possible distributions and durations of anomalies during the next 30 minutes, and we use the following three metrics to evaluate.

- MAE:

$$\begin{cases} \mathrm{Dist}_{\mathrm{dis}}\left(a_{v_i}^{t+1}, \widehat{a_{v_i}^{t+1}}\right) = \sum_t \{\frac{1}{N}\sum_{i=1}^{N} |a_{v_i}^{t+1} - \widehat{a_{v_i}^{t+1}}|\} \\ \mathrm{Dist}_{\mathrm{dur}}\left(d_{v_i}^{t+1}, \widehat{d_{v_i}^{t+1}}\right) = \sum_t \{\frac{1}{N}\sum_{i=1}^{N} |d_{v_i}^{t+1} - \widehat{d_{v_i}^{t+1}}|\} \end{cases} \tag{23}$$

- MSE:

$$\begin{cases} \mathrm{Dist}_{\mathrm{dis}}\left(a_{v_i}^{t+1}, \widehat{a_{v_i}^{t+1}}\right) = \sum_t \{\frac{1}{N}\sum_{i=1}^{N} (a_{v_i}^{t+1} - \widehat{a_{v_i}^{t+1}})^2\} \\ \mathrm{Dist}_{\mathrm{dur}}\left(d_{v_i}^{t+1}, \widehat{d_{v_i}^{t+1}}\right) = \sum_t \{\frac{1}{N}\sum_{i=1}^{N} (d_{v_i}^{t+1} - \widehat{d_{v_i}^{t+1}})^2\} \end{cases} \tag{24}$$

- Accuracy of top K(Acc@K): this metric is the percentage of accurate predictions for a list of predictions with length K [33]. In this paper, we select K highest-risky areas in predicted results and calculate the percentage of accurate predictions in K predicted regions to the total anomalous areas[4]. Considering the actual numbers of anomalous areas in history, we here set $K = 20$ for comparison.

5.3 Baselines

We compare our A2DJP model with the following state-of-the-art methods from various research lines, i.e., classical machine learning algorithms, conventional time series forecasting methods, spatiotemporal forecasting methods, and ensemble learning (regression) models, and all alternative solutions can be categorized into two categories, anomaly distribution prediction methods, and anomaly duration prediction methods. For the sake of fairness, regarding all solutions, we all carry out the predictions from the three above-mentioned temporal granularities. Meanwhile, for all alternative baselines, we first initialize the hyperparameters by referring to their corresponding literature and published codes and then fine-tune the parameters based on the characteristics of our datasets (including but not limited to the number of nodes and number of features). Consequently, we can ensure the fairness of our experiments. The optimal parameters are given after the description of baselines.

5.3.1 *Anomaly distribution prediction methods*

Auto-Regressive Integrated Moving Average (ARIMA) [34]: it is a conventional time series learning model, and is usually used to predict future values in time series. Here the parameter tuple (p, d, q) of ARIMA are set as (1, 2, 6) in both two datasets.

Heterogeneous Convolutional LSTM (Hetero-ConvLSTM) [7]: it is an advanced deep learning method for urban traffic accident prediction. The urban areas of New York City and Chicago are also divided into a 30×22 grid and a 28×19 grid respectively, and the kernel size of CNN are set as 3×3 and 5×5 for New York City and Chicago respectively.

Temporal Graph Convolutional Network(TGCN) [8]: it is a time series traffic flow prediction framework which combines GCN with Gated Recurrent Unit (GRU) network to respectively capture spatial and temporal correlations. Here we set the number of GCN blocks to 2 and the number of GRU hidden units to 64 for both two datasets.

SpatioTemporal GCN (ST-GCN) [35]: it is a multi-step traffic forecasting model which employs several spatiotemporal convolutional to extract both spatial and temporal correlations,

3. Pytorch is an open source machine learning framework that accelerates the path from research prototyping to production deployment.

4. Usually, the number of K is absolutely greater than the total number of anomalous subregions.

and the proposed spatiotemporal convolutional block integrates both the operations of graph convolution and gated temporal convolution. This spatiotemporal learning model can be used to solve our anomaly distribution prediction task, and here each block consists of three layers with 64, 64, 64 filters in New York Dataset, and 32, 32, 32 filters in Chicago Datasets.

Attention based Spatial-Temporal GCN (ASTGCN) [26]: it is a traffic flow prediction method which employs a spatial-temporal attention mechanism to effectively capture the dynamic spatial-temporal correlations within traffic data from the granularities of short-term, middle-term and long-term. We here set the number of input intervals as 8.

Graph Multi-Attention Network(GMAN) [11]: it follows the encoder-decoder architecture where both the encoder and the decoder consist of multiple spatiotemporal attention blocks, and between the encoder and the decoder, a transformer layer is applied to convert encoder embedding to sequence representations. GMAN is designed to achieve traffic flow prediction, and in this paper, we set the number of blocks to 3, the attention heads to 1, and the dimensionality of attention head to 8 for both two datasets.

SpatioTemporal Adaptive Gated GCN (STAG-GCN) [36]: This deep learning method introduces an adaptive graph gating mechanism to dynamically extract selective spatial dependencies within multi-head self-attention mechanism, and correct information deviations caused by artificially defined spatial correlation. In this paper, we use traffic flow similarity matrix instead of semantic neighbor adjacency matrix, and set the number of multi-heads equal to 3. In tcn layers, dilation factors d = 1, 2, 4 and filter size k = 3 in both two datasets.

Riskseqs [37]: This model is a state-of-the-art model for urban anomaly prediction at present. It designs 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. We set connectedness of urban graph to 0.1, interval length to 30min, and number of GCN blocks to 4 in both two datasets.

A2DJP-DIS:We remove the fusion layer, the negative cosine similarity loss component, and the redundant duration loss component from A2DJP to generate A2DJP-DIS which focuses on the single mission of anomaly distribution prediction.

5.3.2 Anomaly duration prediction methods

Linear HP [38]: This method, which incorporates Hawkes process into linear regression network, has strong capability in modeling sequential cascading data.

XGBoost [39]: This method, which introduces first-order and second-order derivatives as well as regularization to traditional loss functions to prevent overfitting, is well suited for the duration regression task. Here we set learning rate to 0.01, max tree depth to 5 in NeW York Dataset, and set learning rate to 0.015, max tree depth to 3 in Chicago Dataset.

LSTM-FC [40]: An Encoder-Decoder framework using fully connected LSTM units. Here we set the number of LSTM layers is equal to 3 in both two datasets.

Neural HP [31]: This work models the streams of discrete events in continuous time by constructing a neural multivariate point process with a novel continuous-time LSTM, hence extracting the complex influences of past events on future events. Here we set layers as 3 in both two datasets.

Bayesian Neural Network (Bayesian NN) [13]: This work, which consists of a cost-sensitive Bayesian network and a weighted K-nearest neighbor model, predicts durations of future accidents based on a given set of historical accident records and the future new accidents.

Deep Fusion-Restricted Boltzmann Machine (DF-RBM) [15]: DF-RBM, which proposes a deep fusion model based on restricted Boltzmann machines for traffic accident duration prediction, can fully mine nonlinear and complex patterns in traffic accident and flow data. Here we trained two stacked RBMs separately, and then we used full connection layers for imformation fusion.

A2DJP-DUR: We remove the fusion layer, the negative cosine similarity loss component, and the redundant distribution loss component from A2DJP to generate A2DJP-DUR which focuses on the single mission of anomaly duration prediction.

5.4 Performance comparison

To better understand the performances of the proposed A2DJP and the benefits from joint predictions of both distribution and duration, we here separately evaluate the two anomaly forecasting tasks by comparing it with different baseline models. The comprehensive performances are illustrated in Table 4.

From the perspective of anomaly distribution forecasting, we can observe that: i) ARIMA takes temporal dependencies into account while deep learning methods can simultaneously decode both spatial and temporal correlations, therefore most deep learning models outperform ARIMA. Further, those methods, which are embedded with the mechanisms of multi-granularity periodic learning and varietal prior knowledge based data enhancing such as STAG-GCN, RiskSeqs and our proposed approach, can significantly outperform other alternative solutions in predicting the distributions of urban anomalies, and this laterally verifies the effectiveness of this mechanisms. ii) Compared with alternative deep learning model, GNN-based models show better performances in the task of urban anomaly distribution prediction. In particular, the performances of ST-GCN, AST-GCN, STAG-GCN, GMAN and RiskSeqs are better than the performances of Hetro-ConvLSTM with all three metrics. iii) Compared with the state-of-the-art solution on anomaly distribution prediction, RiskSeqs, our model can improve the performance by more than 2% in terms of Acc@20, and it can hit more than 58% and 50% of total anomalies respectively within the 20 highest-risky subregions of New York City and Chicago.

Regarding anomaly duration forecasting, some interesting points can be easily concluded: i) our model, A2DJP, can significantly outperform all other alternative solutions in predicting durations of future anomalies with all three different metrics. Considering that our model is the exclusive GNN-based one in all anomaly duration prediction methods, it also indirectly confirms the superiority of the GNN-based model on predicting anomaly durations and capturing spatiotemporal relationships of urban traffics. ii) Neural HP, which modifies the hidden states of neurons within the LSTM network with degenerative HP modeled influences among anomalies, hence involving the cascading correlations among urban anomalies. With the embedded HP, Neural HP significantly outperforms LSTM-FC in terms of all three different metrics in both New York City and Chicago, and this verifies the validity of HP and deep learning integrated network on urban anomaly duration prediction.

Moreover, as can be easily observed in Table 3, RiskSeqs and DF-RBM can respectively outperform A2DJP-DIS and A2DJP-

TABLE 4: Performance comparisons among different methods.

		Anomaly distributions				Anomaly duration	
	Models	MAE	MSE	Acc@20	Models	MAE	MSE
New York City	ARIMA	2.754	2.987	0.1869	–	–	
	Hetro-ConvLSTM	2.488	2.841	0.2236	–	–	
	TGCN	2.377	2.653	0.3561	Classical HP	0.8724	1.4230
	ST-GCN	1.881	2.492	0.4711	XGBoost	0.6437	0.9058
	ASTGCN	1.855	2.120	0.4805	LSTM-FC	0.6624	0.8703
	GMAN	1.774	1.904	0.4423	Neural HP	0.5217	0.7004
	STAG-GCN	1.429	1.578	0.3088	Bayesian NN	0.5872	0.8859
	RiskSeqs	0.927	1.205	0.5642	DF-RBM	0.5190	0.5803
	A2DJP-DIS	1.425	1.622	0.4329	A2DJP-DUR	0.5927	0.6334
	A2DJP (ours)	**0.787**	**1.112**	**0.5863**	**A2DJP(ours)**	**0.4589**	**0.4902**
Chicago	ARIMA	3.497	4.105	0.1773	–	–	
	Hetro-ConvLSTM	3.471	3.678	0.2107	–	–	
	TGCN	2.698	3.015	0.2754	Classical HP	0.8234	1.3765
	ST-GCN	2.387	2.660	0.3429	XGBoost	0.6013	0.7543
	ASTGCN	2.119	2.472	0.3956	LSTM-FC	0.5568	0.5987
	GMAN	1.703	2.014	0.3820	Neural HP	0.5003	0.5299
	STAG-GCN	1.375	1.561	0.3567	Bayesian NN	0.5201	0.5484
	RiskSeqs	1.207	1.349	0.4801	DF-RBM	0.3463	0.3867
	A2DJP-DIS	1.336	1.546	0.3869	A2DJP-DUR	0.4578	0.5015
	A2DJP (ours)	**1.169**	**1.223**	**0.5007**	**A2DJP(ours)**	**0.2729**	**0.3209**

DUR. Such an interesting phenomenon exactly verifies the effectiveness of our joint-prediction framework. In particular, given the fact that A2DJP-DIS and A2DJP-DUR can be viewed as two variants of A2DJP by respectively removing a prediction task from the joint-prediction framework, the inter-task mutual reinforcement characteristic of this joint-prediction framework is then restrained by the single-task prediction nature of A2DJP-DIS and A2DJP-DUR, hence leading to such phenomenon. We will further investigate the effectiveness of each individual component in our proposed model by conducting a series of ablative experiments in subsequent sections.

In summary, extensive main experiments have verified the superiority of our A2DJP approach on both the predictions of anomaly distributions and durations. Even though the main experiments can laterally verify the effectiveness of some purpose-designed mechanisms such as GPKDE, modified GCN, HP-LSTM, and HP-GCN, the distinctive effect of each individual mechanism and the superiority of the joint-prediction framework should be further clarified through extensive ablative studies. To this end, we conduct a series of ablation experiments in the next subsection.

5.5 Ablation experiments

To verify the validity of each individual component as well as the overall framework of A2DJP on addressing specific challenges, in this subsection, extensive ablation experiments are then conducted to demonstrate how detailed implementations of our model exactly contribute to final improvements.

5.5.1 Joint-prediction framework

As discussed, the output of our joint-prediction framework is the distributions of future anomalies with their counterpart durations. To further investigate the effectiveness of the proposed joint-prediction framework, we select and concatenate two individual baselines with the best performance respectively in the predictions of anomaly distributions and durations to fenerate four new models, i.e., Riskseqs+Neural HP, Riskseqs+DF-RBM, STAG-GCN+Neural HP, and STAG-GCN+DF-RBM, and incorporate these four new models with the Information Fusion (IF) component on our A2DJP to make sure the two sub-tasks in these models can work in a joint-prediction manner, and finally compare the performances of these eight temporarily generated networks with the performance of A2DJP in terms of MAE and MSE. The results are demonstrated in Table 5. Notice that the metric of Acc@20 can only be used to evaluate the accuracy of anomaly distribution prediction, and we here omit it in this part.

TABLE 5: Impacts of joint-prediction framework

	New York City/Chicago	
Model	MAE[8]	MSE[8]
RiskSeqs+Neural HP	1.4278/1.6533	1.9240/2.1741
RiskSeqs+DF-RBM	1.5348/1.7423	1.8459/2.1884
STAG-GCN+Neural HP	1.6033/1.6657	1.6985/1.8834
STAG-GCN+DF-RBM	1.7322/1.8054	1.8511/1.9339
RiskSeqs+Neural HP+IF	1.3766/1.5424	1.7749/1.8235
RiskSeqs+DF-RBM+IF	1.4110/1.6087	1.8327/1.9001
STAG-GCN+Neural HP+IF	1.4588/1.5536	1.5398/1.7636
STAG-GCN+DF-RBM+IF	1.7001/1.7564	1.8034/1.9143
A2DJP (ours)	**1.2319/1.4802**	**1.5234/1.6721**

As can be observed, first of all, even though the IF component is incorporated with the four new generated models, our A2DJP method significantly outperforms these four temporarily generated networks in terms of both MAE and MSE with New York City and Chicago, and we think this kind of improvements can be mainly attributed to the employment of the joint-prediction framework. To further investigate the effectiveness of the joint-prediction framework, we continue analyzing the experimental results and discover

8. Considering the temporarily concatenated networks and the joint-prediction framework are intend to address the predictions of both anomaly distributions and durations, we here use the sum of two MAEs respectively in distribution and durations predictions as the MAE of the joint-task. And the MSE of the joint-task is constructed in the same way.

that, the performances of these four IF embedded new models are respectively better than the performances of themselves without IF component, and this completely verifies the effectiveness of the joint-prediction framework. It is worth mentioning that the influences of the employment of ST-HP cannot be ruled out yet. To this end, we then further investigate the impacts of other purpose-designed mechanisms.

5.5.2 Ablative studies of individual components in A2DJP

To fully estimate and understand the implications of individual components, in this subsection, we construct a series of ablative variants by removing or replacing some component within A2DJP. Notice here we also use the MAE and MSE of the joint-task as the main metrics for evaluating all variants, and we also add the Acc@K criteria back for evaluating the predictions of anomaly distributions for a more comprehensive assessment. The constructed variants are listed as follows,

A2D-GPKDE: We omit our GPKDE mechanism during the data pre-processing period in this ablative variant.

A2D-EF: We use all **entangled features** in $\mathcal{F}_{v_i}^t$ directly for joint prediction, this ablative variant replaces all ST-HP embedded components with typical GCN and LSTM modules.

A2D-SM: In this variant, we replace the **traffic volume similarity matrix** in the modified GCN and LSTM integrated network with the fixed distance-based matrix $\mathcal{D}$ as the ablative variant.

A2D-HPGCN: We here replace the anomaly influence decaying matrix Λ_t in HP-GCN with the fixed distance-based matrix $\mathcal{D}$ to construct this ablative variant.

A2D-HPLSTM: In this ablative variant, we remove the temporal decaying channel from HP-LSTM.

A2D-CLS: During training phase, we remove the **negative cosine loss function** from the overall loss function.

A2DJP-GAT: we replace GCN and HP-GCN in A2DJP directly with GAT [23] to generate A2DJP-GAT.

A2DJP-GIN: we replace GCN and HP-GCN in A2DJP directly with GIN [41] to generate A2DJP-GIN.

A2DJP-HPGIN: we replace GCN and HP-GCN in A2DJP respectively with GIN and HP-GIN to generate A2DJP-HPGIN.

TABLE 6: Performances of ablative variants

	New York City/Chicago		
Variant	MAE	MSE	Acc@20
A2D-GPKDE	3.2701/3.0329	3.8025/3.2537	0.1873/0.2734
A2D-EF	1.4713/1.8365	1.8990/1.9012	0.3369/0.3850
A2D-SM	1.2767/1.9187	1.7132/2.1390	0.5734/0.4624
A2D-HPGCN	1.1819/1.5776	1.4911/1.7543	0.5803/0.4904
A2D-HPLSTM	1.2714/1.6754	1.5323/1.8131	0.5677/0.4824
A2D-CLS	1.3270/1.5438	1.6439/2.0173	0.5053/0.4417
A2DJP-GAT	1.4539/1.6002	1.6138/1.8677	0.5054/0.4422
A2DJP-GIN	1.3890/1.5977	1.5433/1.9048	0.5121/0.4576
A2DJP-HPGIN	1.3549/1.5887	1.4465/1.8011	0.4977/0.4589
A2DJP	**1.1328/1.4726**	**1.5135/1.6702**	**0.5851/0.4987**

The performances of all ablative variants are demonstrated in Table 6. As can be easily observed, the unbroken A2DJP significantly outperforms all alternative ablative variants on all three evaluation metrics, and this verifies the validity of each individual purposed-designed mechanism. Further, there exist some interesting points that can be further discussed: i) The newly-designed global prior knowledge-based data enhancement strategy has the most significant impacts on predictions, and it triples and doubles the distribution prediction accuracies respectively in New York City and Chicago. This indicates that sparse issue is the most critical challenge that should be considered in predicting urban anomalies. ii) Compared with A2D-EF, regarding both the predictions of anomaly distributions and durations, our model is significantly better in terms of all three different metrics in both New York City and Chicago, and this indicates that the separate decouple and model of the correlations between urban traffic features and anomalies as well as the cascading correlations among anomalies can indeed contribute to both the predictions of anomaly distributions and durations. iii) With the elimination of spatial decaying matrix in HP-GCN and temporal decaying channel in HP-LSTM, the performances of our method decrease correspondingly in both distribution and duration predictions. It implicates that the consideration of spatiotemporal direct cascading correlations among anomalies is authentically beneficial to our model. iv) Regarding A2DJP-GAT, A2DJP-GIN, and A2DJP-HPGIN, our A2DJP can still outperform all these three new variants in terms of all metrics. The reasons may be that, compared with A2DJP-GAT and A2DJP-GIN, A2DJP can effectively capture the cascading correlations among anomalies, and compare with A2DJP-HPGIN, the carefully designed dynamic traffic similarity matrix of the GCN component in A2DJP is more capable of capturing the direct similarity information of each vertex pair. Your insights on incorporating GAT and GIN with the joint-prediction framework has enlightened us on further improving our work in the future. Due to the limited space of the article, we only added the experimental results into section 5.5.2.

In summary, these ablative experiments establish a quantitative evaluation framework for detecting individual contributions of each purpose-designed component in our model. From these promising results, it can be easily concluded that each well-designed component in A2DJP plays a vital role in our anomaly distribution and duration joint-prediction.

5.6 Impacts of hyper-parameters

To investigate how different hyper-parameters affect the performance of the proposed framework, we show the tuning processes of hyper-parameters on both two datasets.

5.6.1 Impacts of numbers of GCN blocks and LSTMs

In this subsection, we investigate the impacts of numbers of GCN blocks and LSTMs on the final performances of our proposed model. Considering the performances of alternative models are basically consistent in three different metrics with different prediction missions, to this end, we here take the metric of Acc@20 as the main metric for investigating the impacts of numbers of GCN blocks and LSTMs and tuning corresponding parameters. The results are illustrated in Figure 6, and we can easily discover that: i) the performances of A2DJP in New York City and Chicago are consistently optimal while the numbers of GCN blocks and LSTMs are set to 4 and 2 respectively. Specifically, at this point, our proposed model can achieve 58% in Nee York and 49% in Chicago in terms of Acc@20. ii) After this point, with the increase of the numbers of GCN blocks and LSTMs, the performances of our model decline gradually, and the reason may be the model is trapped in the dilemma of over-fitting.

5.6.2 Impacts of weight parameters in loss function

Regarding all adjustable weight parameters in the overall loss function, i.e., Equation (22), to investigate the impacts of all these

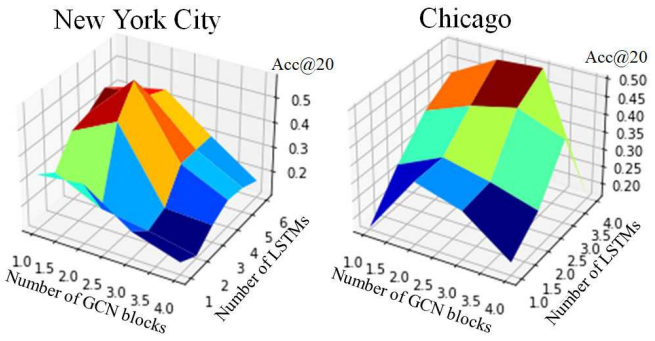

Fig. 6: Impacts of numbers of GCN blocks and LSTMs

parameters, we here set the weight of anomaly distribution loss as 1, and tune the weights of λ_1 and λ_2 by grid searching, and the results are shown in Table 7. As demonstrated, we here selected $\lambda_1 = 1.2$ and $\lambda_2 = 0.8$ as the final setting of the weight parameters in the overall loss function. This parameter combination is also consist with the fact that the absolute error values of distribution predictions is relatively greater the absolute values of duration distribution.

TABLE 7: Hyper-parameter settings in loss function

		New York City/Chicago		
λ_1	λ_2	Acc@20(%)	MAE	MSE
0.8	0.8	56.37/50.12	1.27/1.56	1.54/1.61
0.8	1.2	57.24/48.25	1.34/1.62	1.43/1.69
1.2	0.8	58.51/49.87	1.13/1.47	1.51/1.67
1	1.2	53.15/42.34	1.23/1.52	1.69/1.82
1.2	1	58.07/49.04	1.11/1.12	1.34/1.77
1.5	1	57.14/48.98	1.31/1.69	1.42/1.73
1	1.5	56.23/49.34	1.42/1.78	1.54/1.82

5.6.3 *Impacts of threshold ξ in volume similarity matrix*

To equilibrate the contradiction between the calculation burden and performance of our algorithm, we set a traffic volume similarity threshold ξ to suppress those similarity values in the volume similarity matrices which are less than the threshold. In this subsection, we investigate the impacts of this threshold on the performances of our model and demonstrate the results in Figure 7. From this figure, we observe that the performances of A2DJP are optimal in both New York City and Chicago while the threshold ξ is set to 0.3, and the curves of A2DJP in terms of MAE and MSE haven't change drastically while the threshold increases from 0 to 0.3. With this threshold, more than 15% of total parameters in the volume similarity matrices are suppressed to 0 averagely, which means about 15% of the computational burden are saved while $\xi = 0.3$. For the sake of polytropic equilibrium, ξ is then set to 0.3.

5.6.4 *Impacts of threshold β for filtrating anomalies*

We also set another threshold β to suppress the anomaly numbers and durations of those subregions both to 0 in case that their anomaly numbers are less than the threshold β. To investigate the impacts of this parameter, we then conduct a series of experiments by tuning β from 0 to 1.2 with the step of 0.1, and the results are illustrated in Figure 8. As shown, the

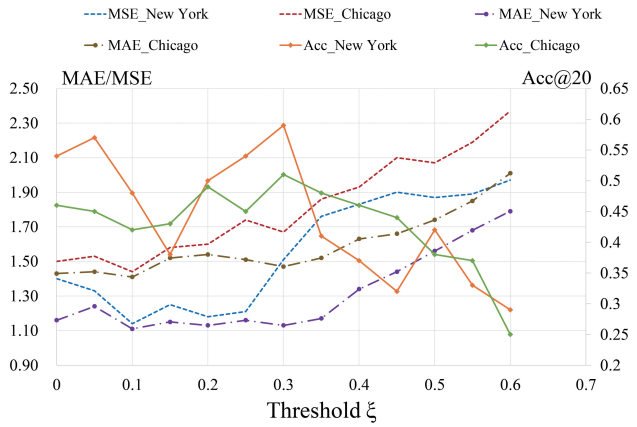

Fig. 7: Performances with different ξ in terms of different metrics on two datasets

performances of our algorithm in terms of Acc@20 increases with the increase of β and achieve the peaks of both two curves while $\beta = 0.8$. Meanwhile, in case that $\beta = 0.8$, the performances of our algorithm in terms of MAE are also optimal for both New York City and Chicago. Therefore, the threshold β is then set to 0.8.

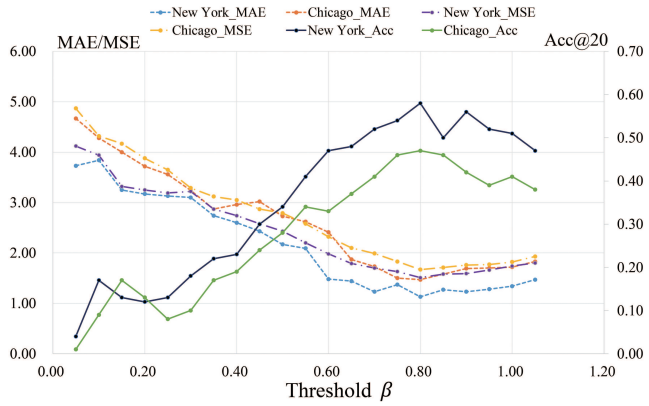

Fig. 8: Performance with different β on two datasets

6 Discussion

In this section, we discuss some interesting issues and lessons learned in this paper.

Case study for analyzing mutual-influences among anomalies: In previous sections, we have introduced and analyzed the self/mutual-exciting characteristics of anomalies, here we demonstrate a case study about these characteristics in Figure 9. First, the influences of two anomalies can be forward superimposed. For instance, the influences of anomaly e_1 decrease with time, and the occurrence of e_2 then enhance the influences, such scenario can be the traffic congestions caused by an accident can be eased gradually with time, and the occurrence of another surrounding accident or traffic restriction may exacerbate the congestions. A similar situation prevails while a new surrounding anomaly, i.e., e_3, e_4, or e_5, happens; Second, the influences of two anomalies can be reversely offset. For instance, the occurrence of e_3 decreases the influences of e_2, and this scenario can be a traffic restriction in the main intersection may cause serious

congestions in this intersection, and an accident in the main road connected to this intersection may decrease the inflow volumes of this intersection, hence remitting the serious congestions in this intersection. Analogously, in case that anomaly e_5 happens, its absolute influences are also remitted by the subsequent influences of e_3. This kind of forward superposition and reverse offset phenomenons among influences of anomalies determine that the cascade correlations among anomalies can be either exciting or suppressing.

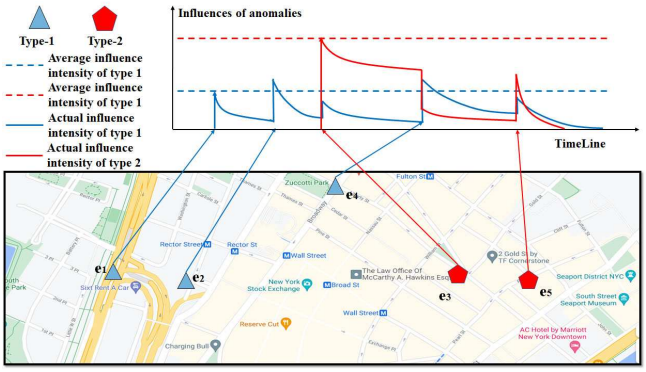

Fig. 9: Case study for analyzing mutual-influences among anomalies: the red and blue dashed lines indicate the average influence intensities of two categories of anomalies, and the corresponding solid lines illustrate the actual influence intensities of corresponding anomalies. The sub-figure in the bottom half illustrates the spatial distribution of five urban anomalies.

The contradiction between performances and long sequence learning: long sequence learning is a critical and challenging issue for deep learnings since it is really hard for LSTM to capture long dependencies while the time span of the input sequence is relatively long. Given the fact that there exist abundant long sequences in inputted traffic data, this may affect the overall performances of our model to a great extent.

Cross-domain generalization of our model: our proposed model can effectively model the spatiotemporal cascade correlations among events that are widespread in the anomalies of different fields such as social media, cloud communication, earthquake, and environmental pollution. Even though temporal HP has been superficially researched in some of the above-mentioned areas, it would be very interesting and necessary to investigate the generalization ability of our proposed model with cross-domain experiments.

7 Conclusion

In this paper, we propose a discrete-continuous task of anomaly prediction with counterpart duration. Compared with traditional anomaly detection methods [6], [7], [37], our framework not only introduces the duration prediction of potential urban events but also proposes to exploit the cascading of targeting events to remedy the feature-oriented event forecasting task. To achieve this, we first decouple the causes of urban anomalies into two categories, i.e., abnormal traffic volumes and cascading influences of previous anomalies. Then we design two GCN-LSTM-based parallel spatiotemporal learning pipelines to capture features-event and event-event correlations, respectively. In particular, inspired by the discrete Hawkes Process, we propose an HP spatiotemporal learning scheme by utilizing and extending the core idea of HP to adaptive learning the propagation and attenuation of event flow influences in spatial and temporal perspectives. Finally, we summarize two categories of causes with weighted fusion and determine potential event duration from two perspectives, risk status, and the spatiotemporal context. To ensure the consistency of the output results, we add cosine similarity loss to the two input vectors to constrain them to keep their consistency in direction. To conclude, our method can enable a more effective and time-aware urban event forecasting with more informative and global traffic statuses. Essentially, our framework provides new insight into discrete-continuous spatiotemporal learning by simultaneously decoupling feature-target disentanglements and determining continuous forecasting by fully leveraging predicted discrete results and various contexts. Last but not least, through effective design, our model enables each module to work effectively and overcomes the shortcomings of previous work.

8 Acknowledgements

This paper is partially supported by the Anhui Science Foundation for Distinguished Young Scholars (No.1908085J24), Natural Science Foundation of China (No.62072427), the Project of Stable Support for Youth Team in Basic Research Field, CAS (No.YSBR005), Natural Science Foundation of Jiangsu Province (No.BK20191193).

References

[1] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1624–1639, 2011.

[2] H. Zhang, Y. Zheng, and Y. Yu, "Detecting urban anomalies using multiple spatio-temporal data sources," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–18, 2018.

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

[4] Q. Chen, X. Song, Z. Fan, T. Xia, H. Yamada, and R. Shibasaki, "A context-aware nonnegative matrix factorization framework for traffic accident risk estimation via heterogeneous data," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2018, pp. 346–351.

[5] B. Sharma, V. K. Katiyar, and K. Kumar, "Traffic accident prediction model using support vector machines with gaussian kernel," in *Proceedings of fifth international conference on soft computing for problem solving*. Springer, 2016, pp. 1–10.

[6] M. Zheng, T. Li, R. Zhu, J. Chen, Z. Ma, M. Tang, Z. Cui, and Z. Wang, "Traffic accident?s severity prediction: A deep-learning approach-based cnn network," *IEEE Access*, vol. 7, pp. 39 897–39 910, 2019.

[7] 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*, 2018, pp. 984–992.

[8] L. Zhao, Y. Song, C. Zhang, Y. Liu, P. Wang, T. Lin, M. Deng, and H. Li, "T-gcn: A temporal graph convolutional network for traffic prediction," *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[9] J. Li, Z. Han, H. Cheng, J. Su, P. Wang, J. Zhang, and L. Pan, "Predicting path failure in time-evolving graphs," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 1279–1289.

[10] 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, no. 01, 2020, pp. 1258–1265.

[11] C. Zheng, X. Fan, C. Wang, and J. Qi, "Gman: A graph multi-attention network for traffic prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 1234–1241.

[12] Q. He, Y. Kamarianakis, K. Jintanakul, and L. Wynter, "Incident duration prediction with hybrid tree-based quantile regression," in *Advances in dynamic network modeling in complex transportation systems*. Springer, 2013, pp. 287–305.
[13] L. Kuang, H. Yan, Y. Zhu, S. Tu, and X. Fan, "Predicting duration of traffic accidents based on cost-sensitive bayesian network and weighted k-nearest neighbor," *Journal of Intelligent Transportation Systems*, vol. 23, no. 2, pp. 161–174, 2019.
[14] L. Shan, Z. Yang, H. Zhang, R. Shi, and L. Kuang, "Predicting duration of traffic accidents based on ensemble learning," in *International Conference on Collaborative Computing: Networking, Applications and Worksharing*. Springer, 2018, pp. 252–266.
[15] L. Li, X. Sheng, B. Du, Y. Wang, and B. Ran, "A deep fusion model based on restricted boltzmann machines for traffic accident duration prediction," *Engineering Applications of Artificial Intelligence*, vol. 93, p. 103686, 2020.
[16] S. D. Tirtha, S. Yasmin, and N. Eluru, "Modeling of incident type and incident duration using data from multiple years," *Analytic Methods in Accident Research*, vol. 28, p. 100132, 2020.
[17] B. Wang, Y. Lin, S. Guo, and H. Wan, "Gsnet: Learning spatial-temporal correlations from geographical and semantic aspects for traffic accident risk forecasting," in *Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI 2021)*, 2021, pp. 4402–4409.
[18] R. W. Heath, M. Kountouris, and T. Bai, "Modeling heterogeneous network interference using poisson point processes," *IEEE Transactions on Signal Processing*, vol. 61, no. 16, pp. 4114–4126, 2013.
[19] F. Ilhan and S. S. Kozat, "Modeling of spatio-temporal hawkes processes with randomized kernels," *IEEE Transactions on Signal Processing*, vol. 68, pp. 4946–4958, 2020.
[20] N. Y. P. Department, "New york opendata," 2020. [Online]. Available: https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95
[21] S. Moosavi, M. H. Samavatian, S. Parthasarathy, and R. Ramnath, "A countrywide traffic accident dataset," *arXiv preprint arXiv:1906.05409*, 2019.
[22] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," *arXiv preprint arXiv:1312.6203*, 2013.
[23] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.
[24] 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.
[25] 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.
[26] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention based spatial-temporal graph convolutional networks for traffic flow forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 922–929.
[27] 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*. PMLR, 2019, pp. 21–29.
[28] H. Hong, H. Guo, Y. Lin, X. Yang, Z. Li, and J. Ye, "An attention-based graph neural network for heterogeneous structural learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 4132–4139.
[29] D. Wang, J. Zhang, W. Cao, J. Li, and Y. Zheng, "When will you arrive? estimating travel time based on deep neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
[30] T. J. Liniger, "Multivariate hawkes processes," Ph.D. dissertation, ETH Zurich, 2009.
[31] H. Mei and J. Eisner, "The neural hawkes process: A neurally self-modulating multivariate point process," *arXiv preprint arXiv:1612.09328*, 2016.
[32] S. Zuo, H. Jiang, Z. Li, T. Zhao, and H. Zha, "Transformer hawkes process," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 692–11 702.
[33] 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.
[34] L. Barba, N. Rodríguez, and C. Montt, "Smoothing strategies combined with arima and neural networks to improve the forecasting of traffic accidents," *The Scientific World Journal*, vol. 2014, 2014.
[35] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.
[36] B. Lu, X. Gan, H. Jin, L. Fu, and H. Zhang, "Spatiotemporal adaptive gated graph convolution network for urban traffic flow forecasting," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 1025–1034.
[37] Z. Zhou, Y. Wang, X. Xie, L. Chen, and C. Zhu, "Foresee urban sparse traffic accidents: A spatiotemporal multi-granularity perspective," *IEEE Transactions on Knowledge and Data Engineering*, 2020.
[38] M. Lukasik, P. Srijith, D. Vu, K. Bontcheva, A. Zubiaga, and T. Cohn, "Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2016, pp. 393–398.
[39] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
[40] M. Sundermeyer, R. Schlüter, and H. Ney, "Lstm neural networks for language modeling," in *Thirteenth annual conference of the international speech communication association*. ISCA, 2012, pp. 194–197.
[41] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" *arXiv preprint arXiv:1810.00826*, 2018.

Kun Wang 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.

Zhengyang Zhou is now a doctoral student 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.

Xu Wang is now a doctoral student in the School of Data Science, University of Science and Technology of China. He got his bachelor degree of automation at North Eastern University in 2017. His research interest mainly includes data mining, machine learning and computer vision

Pengkun Wang is now a doctoral student in the School of Data Science, University of Science and Technology of China. He got his bachelor degree of automation at Jilin University in 2017. His research interest mainly includes computer vision, data mining and machine learning.

Qi Fang is now a master student in the College of Computer Science and Technology , University of Harbin Engineering. His research interest mainly includes computer vision, data mining and machine learning.

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.