IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 7, 2020, accepted January 14, 2020, date of publication February 3, 2020, date of current version February 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2971341

HUAD: Hierarchical Urban Anomaly Detection Based on Spatio-Temporal Data

XIANGJIE KONG[1,2], (Senior Member, IEEE), HAORAN GAO[2], OSAMA ALFARRAJ[3], QICHAO NI[2], CHAOFAN ZHENG[2], AND GUOJIANG SHEN[1]

[1]College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China
[2]School of Software, Dalian University of Technology, Dalian 116620, China
[3]Computer Science Department, Community College, King Saud University, Riyadh 11437, Saudi Arabia

Corresponding author: Osama Alfarraj (oalfarraj@ksu.edu.sa)

This work was supported by the Researchers Supporting Project of King Saud University, Riyadh, Saudi Arabia, under Grant RSP-2019/102.

ABSTRACT Due to the rapid development of communication and sensing technology, a large amount of mobile data is collected so that we can infer the complex movement laws of humans. For cities, some unusual events may endanger public safety. If the early warning of an abnormal event can be issued, it is of great application value to urban construction services. To detect urban anomalies, this paper proposes the *Hierarchical Urban Anomaly Detection* (*HUAD*) framework. The first step in this framework is to build rough anomaly characteristics that need to be calculated by some traffic flow consisted of subway and taxi data. In the second step, the alternative abnormal regions were obtained. Then, the long short-term memory (LSTM) network is used to predict the traffic to get the historical anomaly scores. Following that, the refined anomaly characteristics are generated from adjacent regions, adjacent periods and historical anomalies. The final abnormal regions were detected by One-Class Support Vector Machine (OC-SVM). At last, based on real data sets, we analyzes the traffic flow of the target region and adjacent regions from multiple perspectives in view of the large crowd gathering activities, and the effectiveness of the method is verified.

INDEX TERMS Spatio-temporal data fusion, traffic flow prediction, urban anomaly detection.

I. INTRODUCTION

With the development of communication technology and sensing technology, massive multi-source heterogeneous data are generated from clients, such as vehicle trajectories, social platform data, geographic information system (GIS) data [1], [2], etc. These available data lay a foundation for the anomaly analysis and detection based on the laws of human movement. For metropolis with hundreds or even tens millions of people, the movement of crowd follows complex but stable pattern [3]. Obviously, a serious stampede that occurred during the New Year celebration in Shanghai will bring huge loss of life and property if there is no timely warning and treatment.

For urban anomalies, major incidents, epidemics, serious accidents, environmental disasters, and terrorist attacks all pose great threats to public safety and order [4]. At present, a variety of city data is at our fingertips and provides us with two abilities: one is to learn from history how to correctly respond to threats that have occurred, and the other is to respond to these threats in time and even predict them in advance [5]. Anomalies in urban areas may be caused by accidents, traffic regulations, protests, sports, celebrations, disasters, and other events. Detecting anomalies can help disperse congestion, diagnose accidents, and improve people's travel experience.

If we can detect anomalies in time or even predict anomalies in advance, we can minimize the losses caused by urban anomalies. This will be of great social value [1]. The paper focuses on local anomalous events (such as concerts, large competitions, major traffic accidents, etc.) that occur in a small space, not in the urban area (such as holidays) [6]. Compared to overall urban events, local anomalies are low predictable, and the influencing factors are more complex, these making them more challenging [7].

This paper proposed a new framework to detect on urban anomalies. After sufficient data preprocessing, taxi and subway data reflecting passenger flow at various times in

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Chen.

each area are obtained. In anomaly detection, a series of characteristics are constructed for each time period in each region and used for the input of the OC-SVM algorithm for abnormal detection.

Firstly, anomalies in different regions and different time periods are constructed based on the similarity between different regions and different data sources [1], and the candidate anomaly regions are filtered out by the OC-SVM. On this basis, to construct additional anomaly characteristics, we first adopt Long Short-Term Memory (LSTM) [8] to learn traffic data and predict the traffic flow in the last 4 time periods [9]. In order to strengthen the training effect and improve training accuracy, we have carried out data enhancement and normalization on data to prevent overfitting. The difference between prediction and actual flow is taken as a one-dimensional independent characteristic. In addition, we also calculate the Pearson Correlation Coefficient (PCC) of the traffic sequence of the current adjacent period and the traffic sequence corresponding to the adjacent period one week ago in each region [10], while paying attention to the trend of traffic changes in the adjacent region [11]. In the above, a series of related indicators are aggregated to form the characteristic input of OC-SVM, thereby anomalous events in the anomalous area screened [3]. This process aims to improve the accuracy of the model.

Finally, by visualizing the changes in the traffic area during the relevant time period of the Shanghai Mercedes-Benz Arena, we can clearly see the changes in the traffic flow of the concert area in the target area and adjacent areas.

The contributions of this paper includes:

1) A novel framework named *Hierarchical Urban Anomaly Detection* (*HUAD*) is propose to detect urban anomaly region based on spatio-temporal data.
2) The spatio-temporal data from taxi and subway are integrated to improve the accuracy of the method.
3) For a region, the correlation anomalies of adjacent time periods and adjacent regions as well as the influence of predicted values on the anomaly scores are considered.

The rest of this paper is structured as follows. Section II presents some related studies and Section III describes the framework of urban anomaly region detection. The detailed model is described in Section IV and Section V. Experimental results are introduced in Section VI. Finally, we conclude this work
in Section VII.

II. RELATED WORK

A large amount of heterogeneous network data in cities is collected timely consists in that the sensor networks and communication technologies are developing rapidly. New technologies also make it possible to predict urban anomalous events through spatio-temporal network data. On the basis of the huge practicality of urban anomaly detection, it has been widely studied in recent years [12].

A. SPATIO-TEMPORAL DATA PROCESSING AND APPLICATIONS

With the fast development of various positioning techniques such as Global Position System (GPS), mobile devices and remote sensing, spatio-temporal data has become increasingly available nowadays. This survey [13] summarizes the characteristics of spatio-temporal data, common processing methods and related applications. Gao et al. proposed multimodal deep learning model based on spatio-temporal data witch can handle complex nonlinear urban traffic flow predictions with satisfactory accuracy and effectiveness [14]. By analyzing the spatial and temporal patterns of traffic accident frequency, Ren et al. presented the spatio-temporal correlation of traffic accidents [15]. Sun et al. propose a deep neural network model based on spatio-temporal data to identify non-recurring traffic congestion and explain its causes [16]. Wang et al. develop a two-stage method to effectively detect traffic anomalies from GPS snippets, thus solving the problem of noise and sparsity of GPS fragments collected by vehicles [17].

B. URBAN ANOMALY DETECTION

Recently, artificial neural network models have achieved considerable success in various machine learning tasks [18]. Due to its strong hierarchical feature learning ability in the space-time domain, it has been widely used in urban big data prediction learning, representation learning, anomaly detection and classification. Sudrich et al. discuss graph modeling and the application of anomaly detection for urban data [19]. Based on Bayesian network, a reputation model is proposed by Zhang et al. for the selection of credible sample points to anomaly detection [20]. In order to reveal the characteristics of regional traffic flow patterns in large road networks, He et al. employ dictionary-based compression theory to identify the features of both spatial and temporal patterns by analyzing the multi-dimensional traffic-related data [21]. Zhang et al. perform a two-stage OC-SVM with radial basis function(rbf) kernel to select anomalies [7].

III. PRELIMINARIES

This section introduces the *HUAD* framework for resolving urban anomaly detection. First, the traffic flow matrix is generated by traffic data and the regions. Secondly, the constructed anomaly characteristics I (ACI) is regarded as the input of OC-SVM so that we obtained the anomaly score I. The alternative regions were selected according to the score ranking. Next, LSTM multi-step prediction is used to predict the flow in multiple steps, and the historical anomaly score (HAS) is obtained by comparing it with the real data. Finally, the historical anomaly score is combined with traffic flow data to generate the anomaly characteristics II (ACII) of the alternative regions, and the anomaly score II is calculated by OC-SVM. The regions with high anomaly score II are the final abnormal regions. There are several related definitions aiming to express the framework conveniently.

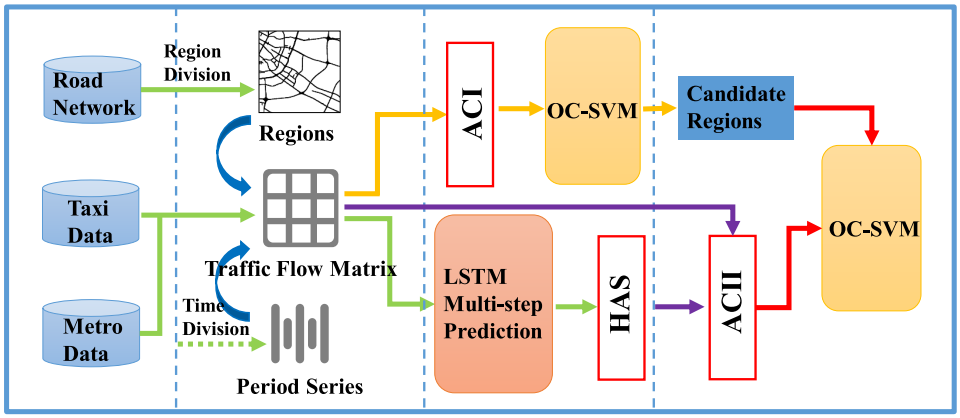


FIGURE 1. The framework of hierarchical urban anomaly detection.

Definition 1 (Region): According to different studies, there are many types of spatial divisions based on various standards and granularities. In this paper, Shanghai is divided into regions through the main road network, and they are denoted as $r = r_1, r_2, \ldots, r_{n_r}$, where n_r is the number of regions. In this paper, n_r is equal to 541.

Definition 2 (Time Period): A day is divided into 48 time periods, each time interval is 30 minutes, denoted as $t = t_1, t_2, \ldots, t_{n_t}$, where n_t is the time period of a month, about $1440(30 * 48)$.

Definition 3 (Data Type): For the raw data of subways and taxis, it is processed into OD data and mapped into a traffic flow matrix. Different traffic flow matrices represent diversified data sources, where the flow to an area over a period of time is defined as two data sources, the inflow and outflow. So we define the category of data $D = D_{taxi_{in}}, D_{taxi_{out}}, D_{metro_{in}}, D_{metro_{out}} = D_1, D_2, D_3, D_4$.

Definition 4 (Traffic Flow Matrix): Define a matrix TR, where the element $V_{t_i,r_j}^{D_m} = TR_m(i, j)$ represents the traffic value for the data source D_m in the r_j-th area of the t_i-th time period.

IV. TRAFFIC FLOW PREDICTION MODEL

A. LONG SHOT-TERM MEMORY NEURAL NETWORK

The long short-term memory network [8] is a variant of RNN capable of learning long-term dependence. LSTM was proposed by Hochreiter and Schmidhuber in 1997 [9] and it was refined and popularized by many scholars in the following studies. LSTM is widely used in many fields, it can handle a variety of problems. The LSTM neural network has three gate structures, called input gate, forgetting gate and output gate. The design of gates is to protect and control unit state. LSTM is generated with a definite purpose to avoid the long-term dependency problem. Fig. 2 shows the repeating module containing four interacting layers in LSTM, X_t is the input to the module, and h_t is the output.

B. TRAFFIC FLOW PREDICTION MODEL BASED ON LSTM

In the previous subsection, four traffic flow matrices TR are obtained, then traffic flow prediction will be performed each

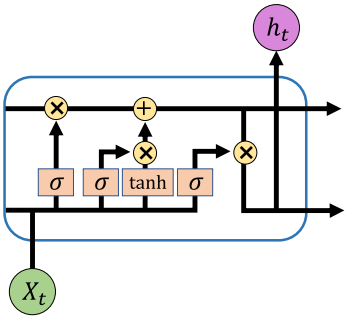


FIGURE 2. Repeating module of LSTM.

traffic flow matrix. First, for a certain region r, the sequence vectors of traffic flow for all time periods are $X = x_1, x_2 \ldots, x_n$. The time prediction model refers to take the known time series X to predict the flow of the following specific time periods $Y = x_{n+1}, x_{n+2} \ldots, x_{n+m}$, and m is the length of the future value to be predicted.

In this paper, in order to carry out abnormal detection for the target time period t and target area r, we first get the time series of the four traffic flow matrices $X = x_1, x_2 \ldots, x_t$. We utilize the data of first 7 days before the target period to predict the flow value 4 periods before the current period. That means the input of LSTM model is $X = x_{t-3-\Delta t+1}, \ldots, x_{t-3}$, the output are flows of four periods before the target period $Y = y_{t-3} \ldots, y_t$. Y's corresponding real value is $X_{real} = x_{t-3}, \ldots\ldots, x_t$, and then calculate the absolute value of the difference between them to get a sequence of differences.

C. MULTI-STEP TIME SERIES PREDICTION MODEL

For LSTM time series prediction, we adopt the sliding window learning method to train and predict the model. The specific training methods are shown in the Fig. 4. In order to rapidly converge the model, we normalized the data set, as (1). Additionally, some simple noises are added to train the model to enhance the robustness of the model.

$$x = \frac{x - min}{max - min} \tag{1}$$

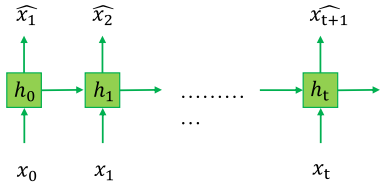


FIGURE 3. Time series training on LSTM.

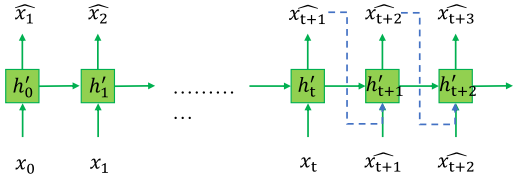


FIGURE 4. LSTM multi-step prediction.

As for a set of sequences X, we input them in batches, such as training sequences of length L, we first determine the length of each training input as len. Given a sequence from x_0 to x_{len-1}, LSTM model will output sequence from $\widehat{x_1}$ to $\widehat{x_{len}}$, its corresponding real sequence is from x_1 to x_{len}. Then the mean square error (MSE) between the predicted value and the real value is loss function showcased as (2). Training keeps MSE as low as possible.

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \widehat{y_i})^2 \tag{2}$$

Inputting the sequence from x_0 to x_{len-1} into trained model M will get sequence from $\widehat{x_1}$ to $\widehat{x_{len}}$. Then we utilize from x_1 to x_{len-1}, appending $\widehat{x_{len}}$ to predict $\widehat{x_{len+1}}$. The results of the multi-step prediction are generated by taking the predicted value to continue the backward prediction. All we need is the predicted sequence value four steps behind the original sequence. As shown in Fig. 4, the value generated by the prediction is used as input to continue the prediction for the next period.

D. HISTORICAL ANOMALY SCORE GENERATION

The four predicted periods are obtained, they are denoted as (3).

$$A = \begin{bmatrix} \widehat{x^1_{D_1}} & \widehat{x^1_{D_2}} & \widehat{x^1_{D_3}} & \widehat{x^1_{D_4}} \\ \widehat{x^2_{D_1}} & \widehat{x^2_{D_2}} & \widehat{x^2_{D_3}} & \widehat{x^2_{D_4}} \\ \widehat{x^3_{D_1}} & \widehat{x^3_{D_2}} & \widehat{x^3_{D_3}} & \widehat{x^3_{D_4}} \\ x^4_{D_1} & x^4_{D_2} & x^4_{D_3} & x^4_{D_4} \end{bmatrix} \tag{3}$$

historical anomaly score are obtained as (4), which correspond to the sum of the absolute value of the difference between the predicted value and the actual flow value of the four traffic flow matrices.

$$C_{pre} = \{Score_{taxi_{in}}, Score_{taxi_{out}}, Score_{metro_{in}}, Score_{metro_{out}}\} \tag{4}$$

V. ANOMALY DETECTION MODEL BASED ON OC-SVM

The historical anomaly score are obtained by calculating the difference between the actual normal flow and the predicted flow from multi-step LSTM. This section introduces the correlation between the target area and the neighboring area's flow changes to explore some other abnormal characteristics as the input of OC-SVM.

A. ONE-CLASS SUPPORT VECTOR MACHINE

Assuming that there's a distribution containing the normal sample, the abnormal samples are outside the normal distribution. OC-SVM attempts to gain the ability to spot anomalies without supervised training [22]. It maps the data to the feature space of the corresponding kernel and tries to find a hyperplane as far away from the origin as possible to determine whether the new input data is normal or not. Different kernel functions can be applied to obtain different hyperplanes corresponding to a variety of nonlinear estimators. To separate the data set from the origin, we solve the following quadratic program (5)(6):

$$\min_{\omega,\xi,\rho} \frac{1}{2}\|\omega\|^2 + \frac{1}{vl}\sum_i \xi_i - \rho \tag{5}$$

$$(\omega \cdot \Phi(X_i)) \geq \rho - \xi_i, \xi_i \geq 0 \tag{6}$$

The parameter v represents the ratio of support vectors to outliers. ξ_i is nonzero slack variable. Φ is the function witch mapping data to high-dimensional feature space.

Therefore, we construct some characteristics related to urban travel traffic flow and apply them to OC-SVM for anomaly detection. The following section will describe the method in detail.

B. INDEPENDENT ABNORMAL SCORE CALCULATION

For each period t, we compute a pairwise similarity matrix $S^t \in R^{r\cdot D^* r\cdot D}$ for each region r and each data source r. $S^t_{r_1,D_1;r_2D_2}$ represents the similarity between the data source D_1 in the region r_1 and the data source D_2 in the region r_2 during the period t. $S^t_{r_1,D_1;r_2D_2}$ can be obtained by (7). Here ρ represents PCC. l is used to control the number of values of interest in calculating similarity [7]. Additionally, we set 1 to cover the length of the whole week:

$$S^t_{r_1,D_1;r_2D_2} = \rho\left(v^{r_1,D_1}_{(t-l+1):t'} v^{r_2,D_2}_{(t-l+1):t}\right) \tag{7}$$

Θ is the threshold, when $S^{t-1}_{r_1,D_1;r_2,D_2} > \theta$, $< r_1, D_1 >$ and $< r_2, D_2 >$ are considered historically similar. When calculating the similarity reduction matrix, we let $SC^t_{r_1,D_1;r_2D_2}$ denote time period $t-1$ and time period t, and the similarity between the data source D_1 in the region r_1 and the data source D_2 in the region r_2 is reduced. It can be calculated by (8). Finally, (9) is to calculate the degree of anomaly $ad^{r,D}_t$ [25] for each $v^{r,D}_t$.

$$SC^t_{r_1,D_1;r_2,D_2} = \max\left(0, S^{t-1}_{r_1,D_1,r_2,D_2} - S^t_{r_1,D_1;r_2,D_2}\right) \tag{8}$$

$$ad_t^{r,D} = \frac{\sum_{<r',D'>\in HS_{r,D}^t} S_{r,D;r',D'}^{t-1}, r^{t-1}, d \cdot SC_{r,D;r',D'}^t}{\sum_{<r',D'>\in HS_{r,D}^t} S_{r,D;r',D'}^{t-1}} \quad (9)$$

C. REGIONAL COMPREHENSIVE ANOMALY SCORE AND ANOMALY DETECTION

This section introduces the aggregation of various abnormal characteristics and anomaly detection by OC-SVM. The calculation of regional anomaly score is divided into two steps. In the first step, we calculate the anomaly score I of the current period in each region. Specifically, for the t period of region r, construct (10) witch is known as ACI:

$$X_{r_t^r} =< ad_t^{r,D_1}, ad_t^{r,D_2}, \ldots, ad_t^{r,D_{ns}} > \quad (10)$$

χ_r is the set with all $X_{r_t^r}$.

OC-SVM with rbf kernel will be trained on χ_r. Giving a new $X_{r_t^r}$, the model will output abnormal points $score_{r_t^r}$. new $X_{r_t^r}$ will be added to the χ_r for further training. $< r, t >$ with the β largest $score_{r_t^r}$ in the past 24 hours are choose as the candidate regions. The definition of candidate regions set C is as follows (11):

$$C = \underset{r,t}{\text{argmax}} [l \cdot n_r \cdot \beta] \{score\_r_t^r | r \in [1, n_r], t \in [T - l + 1, T]\} \quad (11)$$

In the set above, l is the number of periods in a day, and $argmax[k]$ represents the function that returns the index of the maximum k value. In the second step, as well as the nearby region $ad_t^{r,D}$ and the historical anomaly score,the adjacent periods $ad_t^{r,D}$ are considered to calculate the regional anomaly score. Candidate regions with high anomaly scores are final outputs.

Specifically, for the t period of region r, (12) is regard as ACII:

$$\begin{aligned} X_{int_t^r} =& < ad_t^{r,D_1}, ad_t^{r,D_2}, \ldots, ad_t^{r,D_{ns}}, \\ & ad_{t-1}^{r,D_1}, ad_{t-1}^{r,D_2}, \ldots, ad_{t-1}^{r,D_{ns}}, \\ & ad_{t-t_\Delta+1}^{r,D_1}, ad_{t-t_\Delta+1}^{r,D_2}, \ldots, ad_{t-t_\Delta+1}^{r,D_{n_s}}, \\ & ad\_nearby_t^{r,D_1}, \ldots, ad\_nearby_t^{r,D_{ns}}, \\ & Score_{taxi_{in}}, Score_{taxi_{out}}, \\ & Score_{metro_{in}}, Score_{metro_{cut}} > \end{aligned} \quad (12)$$

Line 1 is the same thing as $X_{r_t^r}$. Line 2 to Line 3 are the scores for the region r in the previous $t_\Delta - 1$ time period, where t_Δ represents the number of consecutive periods considered. Line 4 represents the score for the area around r. Similar with χ_r, χ_{int} contains all $X_{int_t^r}$ of the collection, Line 5 indicates historical anomaly score (HAS).

For adjacent areas, Fig. 5 is the histogram of the distance between each region of Shanghai and its nearest five regions. Since the distance amount most areas is about 1200 meters, the adjacent areas are selected with a radius of 1200 meters, and the nearest two areas are taken as the adjacent areas of the target area.

We train OC-SVM on $X_{int_t^r}$, just like in the first stage. Given a new $X_{int_t^r}$, the model will output ASII $score\_int_t^r$.

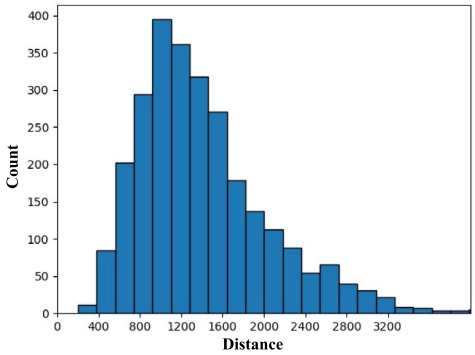


FIGURE 5. Histogram of interregional distance distribution.

Finally, we sort all $< r, t >$ in the candidate set C according to $score\_int_t^r$ and select the $< r, t >$ with the high value of the exception as the region of the exception, as (13):

$$C_{final} = \underset{r,t}{\text{argmax}} [l \cdot n_r \cdot \alpha] \{score\_int_t^r | r \in [1, n_r], t \in [T - l + 1, T], score\_r_t^r \in C\} \quad (13)$$

VI. EVALUATION

This section concentrates on analysis of the experimental results. First, we search some information about large-scale events through the Internet, then treat them as criteria for the accuracy of anomaly detection. In part one, the key links of the experiment are compared by us, while the accuracy of the model is judged recall rate (4). In part two, visual flow analysis of concert was performed to showcase the impact of large-scale events on urban traffic conditions.

The recall rate can be applied to represent the proportion of detected abnormal events to the total, where TP represents the number of detected abnormal events, and FN refers to the model not detected in the abnormal events:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (14)$$

A. DATA PREPROCESSING

As the rapid growth of sensing technology and the information era, the total number of taxis in China has soared in recent years, by one estimate, it's a trend that an average growth rate is 17% annually. The municipal government have imported GPS technology to the taxi industry in order to better manage the planning. The vehicle performs comprehensive monitoring, real-time display of the vehicle's location. The technology can save a large chunk of generated urban traffic data, which is convenient for smart city construction.

Taxi data can directly reflect the status of urban road traffic and passenger flow. The capacity of taxis in metropolis, however, can only embody the changes in actual traffic flow unilaterally. Compared to the capacity of taxis, the capacity of subways with thousands of times of improvement can better reflect the movement of citizens in the city. The subway data is different from the GPS data of taxis. It has clear start and end locations, field information is easier to process, and

TABLE 1. Specific information of taxi GPS trajectory data.

Field	Annotation	The sample
TaxiID	Taxi number	2201252167
Latitude	Coordinates	121.465545
Longitude	Coordinates	31.224068
State	1:passengers, 2:no-passengers	1
date	The date the GPS record was sent	2015-04-25
Time	The specific time the GPS record was sent	13:17:00
speed	Taxi speed	6.613991

TABLE 2. Taxi statistics.

Attribute	Parameter	Detailed description
Time span	April,2015	From 1st to 30th
Number of days	30	21 working days
Number of taxis	13695	Includes 25% Shanghai taxis
Data set size	34 billion items of GPS data	619GB

TABLE 3. Display of subway data.

Field	Annotation	The sample
Card number	One-card-pass number	602141128
Trade date	Trade date	2015-04-1
Trade hour	Trade hour	7:51:8
Station name	Name of the subway station where the card is used	Line 1 xinzhuang station
Amount of money	Charge by credit card	0
Industry name	Mode of transportation (subway, bus)	Subway
Nature of the trading	Preferential or non-preferential	Preferential

the origin-destination (OD) response is more obvious. It is very suitable for identifying abnormal behavior such as large crowds gathering quickly in cities.

1) DATA DESCRIPTION

First, a brief description of the taxi GPS data is shown in Table 1 and Table 2. Each piece of data in this dataset includes 7 fields. The unit of speed is km/h. In this paper, the original GPS data is routinely preprocessed, and some dirty data and invalid data are eliminated. Based on the processed data, the passenger data is extracted as valid data information based on the field time and the vehicle id. For the subway data, as shown in Table 3 and Table 4, the data contains fields such as card number, transaction time, transaction date, chinese name of the subway line, transaction amount, and other fields.

2) SPACE AND TIME DIVISION

Above all, this paper divides Shanghai into regions [23]. The purpose is to form cleanse OD data and form the required traffic flow matrix in the subsequent processing of subway and taxi traffic data. Taking the city's main road network as a framework, we divided Shanghai into 541 regions, and the whole urban area in this study was between N31.15-N31.37 and E121.31-E121.84, as shown in Fig. 6. These areas do not overlap with each other, and each area is naturally divided by the urban road. Accordingly, dividing the urban area by roads. is more natural, and it

TABLE 4. Subway statistics.

Attribute	Parameter	Detailed description
Time span	April,2015	From 1st to 30th
Number of days	30	21 working days
Number of taxis	13695	Includes 25% Shanghai taxis
Data set size	445 million pieces of data	24GB

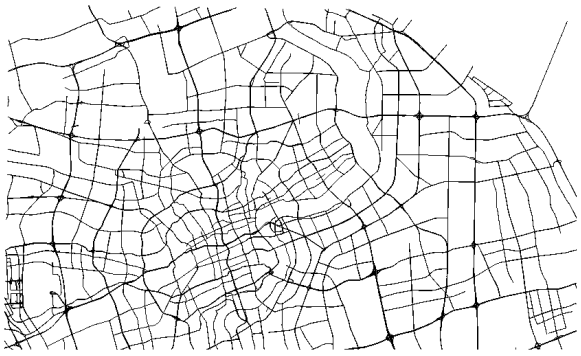


FIGURE 6. Regional division of Shanghai.

also has a certain abstract modeling ability. Recording these areas $R = r_1, r_2, \ldots, r_{541}$.

In order to comply with the prediction input of the time series model, we consider such conditions to more clearly reflect the changes in the traffic flow in phases. The 24 hours a day is divided into 48 periods, and each period is 30 minutes in length to obtain a time series $T = t_1, t_2, \ldots, t_{1440}$.

3) TRAFFIC FLOW MATRIX GENERATION

For taxi data, after pre-processing, the data now available is a series of taxi status information. We classify each taxi according to GPS time and determine the passenger's boarding location based on the loading status $iloc_{st}$, location for alighting $iloc_{ed}$, boarding time st, alighting time ed. Then we get the area number corresponding to the boarding point and the alighting point r_1, r_2 by the latitude and longitude, the serial number of the time period corresponding to the boarding time and the boarding time t_1, t_2. Following that, we determine the value $V_{t_1,r_1}^{D_{taxi_{in}}}$ and $V_{t_2,r_2}^{D_{taxi_{out}}}$ add 1 to each value to get a matrix of two $n_t * n_r(1440 * 541)$ taxi data.

For subway data, after removing invalid data, we notice that most remaining data is effectively regular. Initially, we classify the data of the day according to the card number id and time, and then determine whether to get on or off, area number r_1, r_2, and boarding time t_1, t_2 according to the consumption amount of the card. Then we determine the value $V_{t_1,r_1}^{D_{metro_{in}}}$ and $V_{t_1,r_1}^{D_{metro_{out}}}$ adding 1, meanwhile, we delete the original two data at the same time. Since the operating time of the Shanghai Metro is more than 00:00 on the day, we have deleted the source data after extracting valid information so that it can avoid missing data. The remaining data is divided into two parts: one is invalid data, the other is passengers who get on the vehicle on a day, but get off on the next day. Hence, as for the second type of data, we add it to the next day when

TABLE 5. April 2015 large-scale event in Shanghai.

Event type	Address	Date	Time Frame
Concert	Shanghai Gymnasium	1st, April	19:30-21:30
Concert	Mercedes-Benz Arena	3rd ,April	19:30-22:00
Concert	Mercedes-Benz Arena	4th, April	19:30-22:00
Football Match	Mercedes-Benz Arena	7th, April	20:00-22:00
Career Fair	Shanghai Stadium	11th, April	09:00-14:00
Concert	Mercedes-Benz Arena	11th, April	19:30-23:00
Football Match	Shanghai Stadium	12th, April	08:00-23:00
Concert	Mercedes-Benz Arena	12th, April	19:30-23:00
Concert	Mercedes-Benz Arena	18th, April	19:30-22:00
Football Match	Shanghai Stadium	19th, April	08:00-23:00
Concert	Mercedes-Benz Arena	21th, April	19:30-21:30
Concert	Mercedes-Benz Arena	22th, April	19:30-21:30
Concert	Shanghai Centre Theatre	26th, April	20:00-22:00
Concert	Shanghai Gymnasium	28th, April	19:30-22:00

processing data, and match the traffic flow matrix information to make full use of the data.

After the data preprocessing, we generated four corresponding traffic flow matrices *TR*, namely, taxi inflow traffic flow matrix, taxi outflow traffic flow matrix, subway inflow traffic flow matrix, and subway outflow traffic flow matrix.

B. RESULTS ANALYSIS

1) LIST OF LARGE CROWD GATHERING ACTIVITIES IN SHANGHAI IN APRIL

As shown in Table 5, this paper finds large-scale crowd gathering activities that occurred in Shanghai in April 2015, including megastar concerts and some major football tournaments, which affects the traffic conditions in this area and surrounding areas.

2) RECALL RATE ANALYSIS AND COMPARISON

First, the location r where the large-scale gathering occurs is given, then we calculate the corresponding area number in order to get the corresponding period number t according to the start and end time. Mark the tag corresponding to the $< r, t >$ tuple as 1, other periods are marked as 0. After marking the abnormal data, we perform anomaly detection according to the method described in the previous section and analyze the experimental results corresponding to real data. Because the anomaly information obtained in this article is relatively limited, while the entire city is promoting development and traffic congestion anomalies may occur due to some complex and unknown factors. In addition, through data observation and analysis, compared with some large-scale activities and performances, the morning and evening peaks in cities have a greater impact on urban traffic. Therefore, analyzing the local area anomalies caused by these large-scale performances, we apply the recall rate to measure the accuracy of the model judgment. Recall rate is calculated by (14).

In order to show the effect of the model, we split and simplify the two key steps of the model, conduct experiments for comparison, and analyze key indicators such as multiple data sources, time series prediction indicators, adjacent areas, and adjacent periods. Table 6 demonstrates the experimental results of key components in the model.

TABLE 6. Experimental outcomes and comparison.

Method	Describe	Recall
$X_r + X_{int}(\Delta t, nearby, LSTM)$	Use taxi and subway data, including adjacent time periods, adjacent areas, and LSTM predictive characteristics	0.71
$X_r + X_{int}(\Delta t, nearby, LSTM)_{taxi}$	Use taxi data, including adjacent time periods, adjacent areas, and LSTM predictive characteristics	0.29
$X_r + X_{int}(\Delta t, nearby, LSTM)_{metro}$	Use subway data, including adjacent time periods, adjacent areas, and LSTM predictive characteristics	0.50
$X_r + X_{int}(\Delta t, nearby, ARIMA)$	Use taxi and subway data, including adjacent time periods, adjacent areas, and ARIMA predictive characteristics	0.57
$X_r + X_{int}(\Delta t, nearby)$	Use taxi and subway data, including adjacent time periods, adjacent areas	0.42
$X_r + X_{int}(nearby, LSTM)$	Use taxi and subway data, including adjacent areas, and LSTM predictive characteristics	0.36
$X_r + X_{int}(\Delta t, LSTM)$	Use taxi and subway data, including adjacent time periods, and LSTM predictive characteristics	0.43

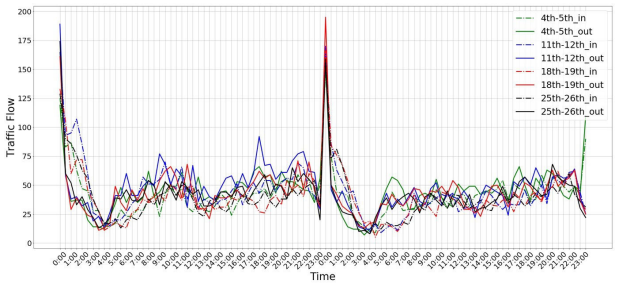


FIGURE 7. Taxi traffic flow in area 32 every Saturday and Sunday in April.

Conversely, the Fig. 8 shows the trend of passenger flow in China Art Museum station on Saturday and Sunday in April. China Art Museum station is not only the main passenger flow station in the target area 32, but is the nearest subway station to Mercedes-Benz Arena. The flow trend of this station can directly reflect the flow of subway passengers in this area to some extent. First of all, from Fig. 8, it can be seen that the subway capacity is more than ten times that of a taxi, and the gap between subway and taxi capacity is more obvious in downtown areas. Secondly, the interval from 16:00 to 23:00 on Saturday and Sunday drawn with dotted boxes in Fig. 8 is particularly visualized in Fig. 9. According to Fig. 9, it can be observed that the passenger flow on April 11th, 12th, 18th at 6 p.m. and around 10 p.m.

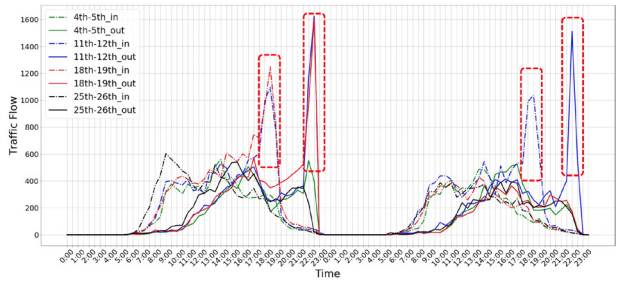


FIGURE 8. Traffic flow of China Art Museum station on Saturday and Sunday in April.

increases significantly compared with the normal weekend. And these three periods correspond to three concerts. With the surge of traffic flow, there is a large flow trend gap between the target area and similar areas. The calculated anomalies will also be remarkably increased, which can provide highly useful information for detecting abnormal travel behavior in cities.

C. CASE STUDY

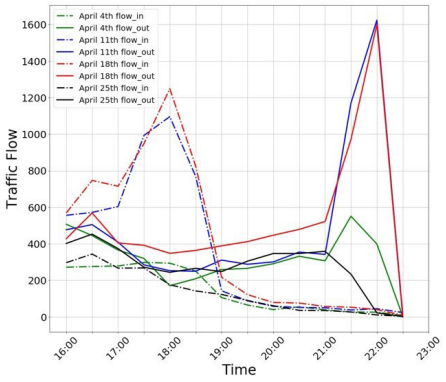
A large crowd gathering event in Shanghai in April 2015 is presented above. This paper devotes to conducting flow analysis and visualization of large-scale concerts given below. Then the key components of the model will be analyzed from multiple perspectives. All the concerts collected in this paper were held in Shanghai's Mercedes-Benz Arena, located in area 32, which includes three subway stations. Among these stations, the nearest subway station to the cultural center is China Art Museum station, which is also the main stop to take the subway to the cultural center.

1) TRAFFIC FLOW ANALYSIS OF TARGET AREA

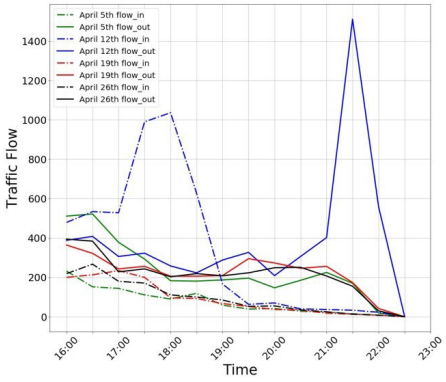
In Fig. 7, it reflects the changing trend of taxi flow on Saturday and Sunday in April, where the dotted line represents the inflow traffic, the solid line represents the outflow traffic, the first 24 hours represent Saturday, and the second 24 hours represent Sunday. Because of the limited capacity of taxi, it can be seen from the Fig. 7 that the taxi flow changes in the area 32 basically keep changing with traffic laws in most cases. Even during the Qingming Festival and in the period of the large-scale concerts, the fluctuation of taxi traffic is not obvious. Due to the concert held by a famous singer, however, the traffic increases slightly at around 18:00 on April 11 and the overall periodicity is obvious. The alteration with large-scale activities is relatively limited. Therefore, the recall rate is low when only the experiment from taxi data is conducted.

2) ANALYSIS OF TRAFFIC FLOW IN ADJACENT AREAS

From the Fig. 10, it demonstrates the Saturday and daily passenger flow of South Xizang Road station in area 59 in April. Area 59 where the exclusively one with a subway station is located is the adjacent area of area 32. It can be concluded from the dotted boxes in Fig. 10, a whopping increase of traffic flow in area 59 occurs when some concerts



(a) Saturday



(b) Sunday

FIGURE 9. Traffic flow of China art museum station at fixed period on Saturday and Sunday.

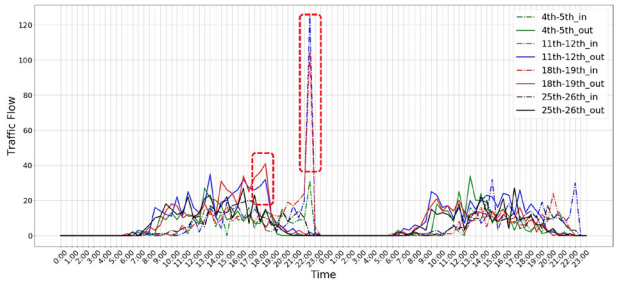


FIGURE 10. Traffic flow of South XiZang Road station, area 59.

have been held. In Fig. 11, the flow visualization is made for the fixed period from 16:00 to 23:00 on all Saturdays in April. Accordingly, the Fig. 10 reflects that at 6:00 p.m. and around 10:30 p.m. on 11th and 18th, the peak section appeared opposite to the traffic flow of China Art Museum station in area 32.

The inflow and outflow surged at 18:00 and 22:00, because China Art Museum station in area 32 was close to the concert venue. On the contrary, for South Xizang Road station in area 59, the outflow increased slightly at 18:00 and the inflow surged at around 22:00. As the regional traffic flow characteristics around the target area, not all concert will affect the adjacent area, such as the concert on April 12. Our model shows the traffic impact caused by large activities improves the recall rate by considering the flow characteristics of the adjacent area.

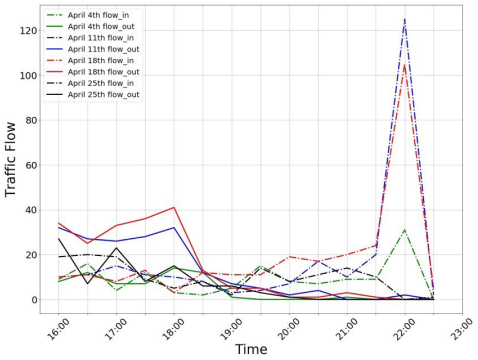


FIGURE 11. Traffic flow of South XiZang Road station at fixed period on Saturday.

According to the variation of subway flow in Fig. 8 and Fig. 9, it can be seen that the surge of passenger flow also experiences a period of growth time. When the traffic peak peaked at 18:00, the passenger flow in the first two periods is constantly ascending. Therefore, the anomalies of the adjacent periods of the target period are also considered in the model.

VII. CONCLUSION

In this paper, *HUAD* framework is proposed to detect regional anomaly. First, we generate regions and time periods sequences, and build a traffic flow matrix based on taxi and subway data. Then historical anomaly scores are obtained by comparing real data with predicted data from multi-step prediction LSTM. In the rough anomaly detection, we construct anomaly characteristics I and get anomaly score I through OC-SVM. Then we pick the candidate region from the outcomes. Next, anomaly characteristics II, including historical anomaly scores, will be put in OC-SVM and tested once more to get final anomaly region. Ultimately, based upon the multi-source data of taxi and subway data in real world, this paper analyzes the traffic flow of the target area and adjacent areas from different perspectives in view of the large crowd gathering activities to verify the validity of the model.

REFERENCES

[1] Z. Ning, Y. Feng, M. Collotta, X. Kong, X. Wang, L. Guo, X. Hu, and B. Hu, "Deep learning in edge of vehicles: Exploring trirelationship for data transmission," *IEEE Trans. Ind. Informat.*, vol. 15, no. 10, pp. 5737–5746, Oct. 2019.

[2] Z. Ning, P. Dong, X. Wang, M. S. Obaidat, X. Hu, L. Guo, Y. Guo, J. Huang, B. Hu, and Y. Li, "When deep reinforcement learning meets 5G vehicular networks: A distributed offloading framework for traffic big data," *IEEE Trans. Ind. Informat.*, vol. 16, no. 2, pp. 1352–1361, Feb. 2020.

[3] X. Kong, X. Liu, B. Jedari, M. Li, L. Wan, and F. Xia, "Mobile crowdsourcing in smart cities: Technologies, applications, and future challenges," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8095–8113, Oct. 2019.

[4] X. Wang, Z. Ning, X. Hu, L. Wang, L. Guo, B. Hu, and X. Wu, "Future communications and energy management in the internet of vehicles: Toward intelligent energy-harvesting," *IEEE Wireless Commun.*, vol. 26, no. 6, pp. 87–93, Dec. 2019.

[5] C. Chen, Y. Ding, X. Xie, S. Zhang, Z. Wang, and L. Feng, "TrajCompressor: An online map-matching-based trajectory compression framework leveraging vehicle heading direction and change," *IEEE Trans. Intell. Transp. Syst.*, to be published.

[6] C. Chen, D. Zhang, X. Ma, B. Guo, L. Wang, Y. Wang, and E. Sha, "Crowddeliver: Planning city-wide package delivery paths leveraging the crowd of taxis," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 6, pp. 1478–1496, Jun. 2017.

[7] H. Zhang, Y. Zheng, and Y. Yu, "Detecting urban anomalies using multiple spatio-temporal data sources," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 1–18, Mar. 2018.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[9] X. Kong, F. Xia, J. Li, M. Hou, M. Li, and Y. Xiang, "A Shared bus profiling scheme for smart cities based on heterogeneous mobile crowdsourced data," *IEEE Trans. Ind. Informat.*, vol. 16, no. 2, pp. 1436–1444, Feb. 2020.

[10] Z. Ning, P. Dong, X. Wang, J. J. Rodrigues, and F. Xia, "Deep reinforcement learning for vehicular edge computing: An intelligent offloading system," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 6, p. 60, 2019.

[11] G. Shen, L. Zhu, J. Lou, S. Shen, Z. Liu, and L. Tang, "Infrared multi-pedestrian tracking in vertical view via siamese convolution network," *IEEE Access*, vol. 7, pp. 42718–42725, 2019.

[12] Z. Ning, J. Huang, X. Wang, J. J. P. C. Rodrigues, and L. Guo, "Mobile edge computing-enabled internet of vehicles: Toward energy-efficient scheduling," *IEEE Netw.*, vol. 33, no. 5, pp. 198–205, Sep. 2019.

[13] S. Wang, J. Cao, and P. S. Yu, "Deep learning for spatio-temporal data mining: A survey," 2019, *arXiv:1906.04928*. [Online]. Available: https://arxiv.org/abs/1906.04928

[14] S. Du, T. Li, X. Gong, Z. Yu, Y. Huang, and S.-J. Horng, "A hybrid method for traffic flow forecasting using multimodal deep learning," 2018, *arXiv:1803.02099*. [Online]. Available: https://arxiv.org/abs/1803.02099

[15] H. Ren, Y. Song, J. Wang, Y. Hu, and J. Lei, "A deep learning approach to the citywide traffic accident risk prediction," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 3346–3351.

[16] F. Sun, A. Dubey, and J. White, "DxNAT—Deep neural networks for explaining non-recurring traffic congestion," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 2141–2150.

[17] W. Kuang, S. An, and H. Jiang, "Detecting traffic anomalies in urban areas using taxi GPS data," *Math. Problems Eng.*, vol. 2015, Sep. 2015, Art. no. 809582.

[18] C. Chen, S. Jiao, S. Zhang, W. Liu, L. Feng, and Y. Wang, "TripImputor: Real-time imputing taxi trip purpose leveraging multi-sourced urban data," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 10, pp. 3292–3304, Oct. 2018.

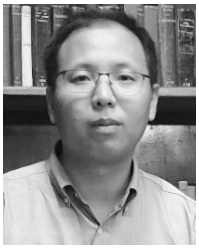
[19] S. Sudrich, J. Borges, and M. Beigl, "Graph-based anomaly detection for smart cities: A survey," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Aug. 2017, pp. 1–7.

[20] H. Zhang and Z. Li, "Anomaly detection approach for urban sensing based on credibility and time-series analysis optimization model," *IEEE Access*, vol. 7, pp. 49102–49110, 2019.

[21] M. He, S. Pathak, U. Muaz, J. Zhou, S. Saini, S. Malinchik, and S. Sobolevsky, "Pattern and anomaly detection in urban temporal networks," 2019, *arXiv:1912.01960*. [Online]. Available: https://arxiv.org/abs/1912.01960

[22] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt, "Support vector method for novelty detection," in *Adv. Neural Inf. Process. Syst.*, 2000, pp. 582–588.

[23] N. J. Yuan, Y. Zheng, and X. Xie, "Segmentation of urban areas using road networks," Microsoft Res. Asia, Beijing, China, Tech. Rep. MSR-TR-2012–65, 2012.



XIANGJIE KONG (Senior Member, IEEE) received the B.Sc. and Ph.D. degrees from Zhejiang University, Hangzhou, China. He is currently a Professor with the College of Computer Science and Technology, Zhejiang University of Technology. He has published over 100 scientific articles in international journals and conferences (with over 70 indexed by ISI SCIE). His research interests include network science, data science, and computational social science. He is a member of ACM and a Senior Member of the CCF.

HAORAN GAO received the B.Sc. degree in cyber engineering from the Dalian University of Technology, China, in 2019, where he is currently pursuing the master's degree with the School of Software. His research interests include urban data mining, network embedding, and data set completion.

OSAMA ALFARRAJ received the master's and Ph.D. degrees in information and communication technology (ICT) from Griffith University, in 2008 and 2013, respectively. His doctoral dissertation investigates the factors influencing the development of Government in Saudi Arabia, and it is a qualitative investigation of the developers' perspectives. He is currently an Associate Professor with ICT, King Saud University, Riyadh, Saudi Arabia. His research interests include electronic commerce, M-government, the Internet of Things, cloud computing, AI, and big data analytics.

QICHAO NI received the B.Sc. degree in automation (program for excellent engineers) from Yanshan University, China, in 2019. He is currently pursuing the master's degree with the School of Software, Dalian University of Technology, China. His research interests include urban data mining, network embedding, and multiple heterogeneous data fusion.

CHAOFAN ZHENG was born in Xuchang, Henan, in 1997. He received the B.E. degree from the Dalian University of Technology, Dalian, China, in 2019. He is currently pursuing the master's degree with the School of Software and Microelectronics, Peking University, China. He is currently pursuing the master's degree with the School of Software and Microelectronics, Peking University, China. His research interests include machine learning and deep learning for big data of urban transportation.

GUOJIANG SHEN received the B.Sc. degree in control theory and control engineering, and the Ph.D. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 1999 and 2004, respectively. He is currently a Professor with the College of Computer Science and Technology, Zhejiang University of Technology. His current research interests include artificial intelligence theory, big data analytics, and intelligent transportation systems.