

Research Article

A Higher-Order Motif-Based Spatiotemporal Graph Imputation Approach for Transportation Networks

Difeng Zhu,^{1,2} Guojiang Shen ,¹ Jingjing Chen ,³ Wenfeng Zhou,¹ and Xiangjie Kong ¹

¹College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China

²Yuanpei College, Shaoxing University, Shaoxing 312000, China

³School of Economics, Fudan University, Shanghai 200433, China

Correspondence should be addressed to Guojiang Shen; gjshen1975@zjut.edu.cn and Xiangjie Kong; xjkong@ieee.org

Received 7 September 2021; Revised 26 December 2021; Accepted 29 December 2021; Published 17 January 2022

Academic Editor: Yan Huang

Copyright © 2022 Difeng Zhu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Due to the incomplete coverage and failure of traffic data collectors during the collection, traffic data usually suffers from information missing. Achieving accurate imputation is critical to the operation of transportation networks. Existing approaches usually focus on the characteristic analysis of temporal variation and adjacent spatial representation, and the consideration of higher-order spatial correlations and continuous data missing attracts more attentions from the academia and industry. In this paper, by leveraging motif-based graph aggregation, we propose a spatiotemporal imputation approach to address the issue of traffic data missing. First, through motif discovery, the higher-order graph aggregation model was presented in traffic networks. It utilized graph convolution network (GCN) to polymerize the correlated segment attributes of the missing data segments. Then, the multitime dimension imputation model based on bidirectional long short-term memory (Bi-LSTM) incorporated the recent, daily-periodic, and weekly-periodic dependencies of the historical data. Finally, the spatial aggregated values and the temporal fusion values were integrated to obtain the results. We conducted comprehensive experiments based on the real-world dataset and discussed the case of random and continuous data missing by different time intervals, and the results showed that the proposed approach was feasible and accurate.

1. Introduction

With the rapid growth of urbanization, intelligent transportation systems (ITS) are widely adopted for the urban management and traffic control [1, 2]. ITS rely on the availability of traffic data to evaluate traffic status and system performance. Assuming that traffic information would be collected in an all-round way, urban commuters can adapt the traffic conditions of urban roads, grasp the law of traffic flow changes, and subsequently promote the development of urban transportation.

In recent years, emerging information technologies, such as fifth-generation networks [3] and edge computing [4], have brought a bit convenience to traffic data collection, and the collected data is usually mobile, multisource, and real time. Unfortunately, due to the frequent occurrence of various types of failures (e.g., power malfunction, device maintenance, and network issues), collected data always

are incomplete [5]. Moreover, due to the high cost of construction and maintenance, the equipment is difficult to cover the entire traffic network [6]. So, the loss of traffic data in the process of data collection is inevitable.

The problem of information missing significantly weakens the data quality, limits the study of transportation networks (e.g., traffic management, urban planning, and route choice), and in worse, may result in false decisions [6–8]. Thus, handling missing values is a premise for traffic data mining and analysis [9]. Making accurate imputation becomes an important research topic in ITS [10].

The key of data imputation is to discover the hidden spatiotemporal information with regard to the neighbouring data [11]. For instance, as shown in Figure 1, it makes use of the spatiotemporal neighbouring values of the missing data, thereby improving the accuracy of data imputation. Bae et al. proposed two cokriging methods that exploited the existence of spatiotemporal dependency in traffic data,

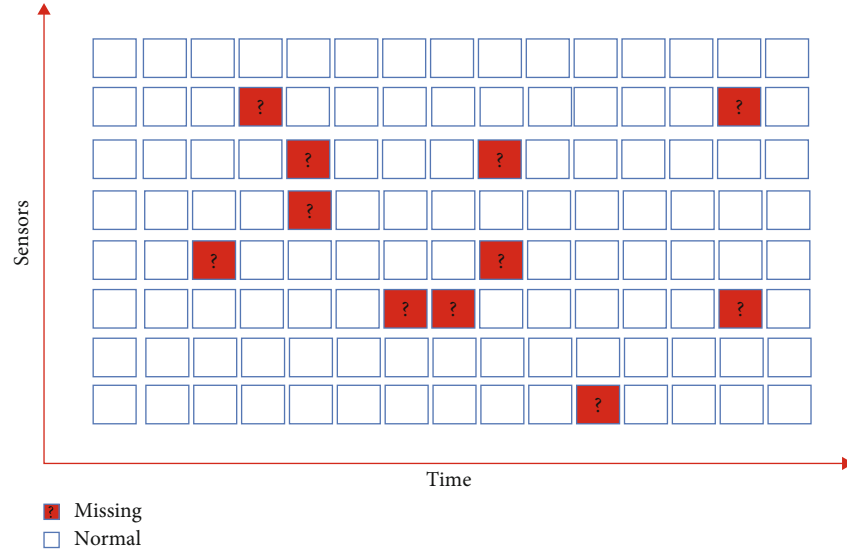


FIGURE 1: Spatiotemporal representation of missing data. The data labelled in red with question mark and the data labelled in white denote the missing values and normal values, respectively.

to impute high-resolution traffic speed under different random data missing scenarios [12]. Li et al. developed a combined deep neural model, which extracted spatio-temporal features to estimate missing values [13]. To characterize the hidden patterns in spatiotemporal traffic data, Chen et al. incorporated a low-rank tensor completion (LRTC) framework with the truncated nuclear norm (TNN) and obtained a better solution for data imputation [14]. Considering the case of continuous data missing, Zhang et al. utilized the temporal neighbouring values of a given period and employed the long short-term memory network (LSTM) to recover missing data [15].

Although many methods achieved promising imputation accuracy of missing data, there are still some limitations. The first question needs to be addressed is how to effectively capture the spatial dependencies. Existing methods usually consider the direct adjacent road segments (i.e., upstream and downstream) but ignore the global information. It is considerable to introduce the imputed models with sensing the global and local variations of spatial information [10]. The second one is the continuous data missing, in which the missing values at some consecutive timestamps in a road segment. In this circumstance, it is unable to generate data during a period of time and provide stable inputs for a model.

To address these issues, in this paper, we propose a spatiotemporal imputation approach for traffic data via motif-based graph aggregation (named MGIA), which incorporates the motif-based spatial aggregation with the multitime dimension fusion by bidirectional LSTM (Bi-LSTM). To the best of our knowledge, our work is the first attempt to apply motif-based spatial method to address the issues of traffic data imputation. The contributions of this paper are summarized as follows:

- (1) We propose a higher-order graph aggregation model based on motifs. It polymerizes the correlated seg-

ment attributes of the missing data segments to capture the higher-order spatial correlations in a road network, which utilizes the method of motif-based search, and the aggregation based on graph convolution network (GCN)

- (2) We develop a Bi-LSTM approach based on the multitime dimension fusion to improve the accuracy in the case of continuous data missing. It incorporates the recent, daily-periodic, and weekly-periodic dependencies to ensure that there are enough historical data to complete the temporal imputation
- (3) We perform extensive experiments on the real-world dataset to evaluate the performance of our approach. The experimental results confirm the advantages of proposed approach with various missing patterns over the state-of-the-art imputation approaches

The remainder of the paper is organised as follows: Section 2 introduces the related work in traffic data imputation. Section 3 describes the proposed approach. Section 4 shows the experimental results and analysis. Section 5 concludes this paper.

2. Related Work

In this section, we introduce the related work regarding the approaches for conventional and deep learning-based imputation.

2.1. Conventional Imputation. In the past decades, traffic data imputation has caused widespread concern. The imputation methods mainly include prediction, interpolation, and statistical methods.

The predictive methods utilize historical data to predict the missing values. The typical methods include Bayesian networks [16] and support vector regression [17]. Ahn

TABLE 1: The limitations of these existing methods.

Category	Limitation
Predictive methods	These methods fail to consider the case of continuous data missing.
Interpolation methods	They focus on the average calculation without considering data global attributes.
Statistical methods	The accuracy of these methods degrades owing to the unknown values of data distribution.
Adversarial learning	Most methods ignore global spatial dependencies in traffic network.
Graph-based methods	They only capture global spatial information from neighbouring data.

et al. then developed a traffic flow prediction method based on Bayesian classifier and support vector regression [18]. It further improved the imputation accuracy of missing values. The hybrid approach based on fuzzy C -means (FCM) is another example of such predictive methods, which integrates the optimized FCM parameters and genetic algorithms to build prediction models [19, 20]. These methods focus more on the historical traffic data for missing data filling and fail to consider the imputation on missing continuous data. Moreover, they ignore the spatial relationships which also provide crucial information for imputation.

Interpolation methods use the average value of the neighbouring data or historical data to impute the missing values. Using traffic data from the same sensor during the same period in neighbouring days, Yin et al. took the average value of these known data to impute the missing values [21]. Chang et al. utilized the k -nearest neighbours (KNN) and local least squares to consider the relationship between similar traffic flow patterns and enhanced the interpolation effects [22]. Kriging interpolation [12, 23] focused on determining the weighted historical values. It considered the spatiotemporal information to capture the characteristics of traffic data. Although these methods can achieve promising imputation results in a short time, they can only focus on the average calculation without considering complicated changes caused by other factors such as data global attributes and random events.

The statistical methods are aimed at developing a data distribution that best fits the imputed missing data. To reflect the uncertainty between the imputation parameters, Audigier et al. designed a multiple imputation method based on Bayesian principal component analysis (BPCA) to cope with incomplete continuous data [24]. To exploit the spatiotemporal correlation of traffic network data, Wang et al. developed a low-rank matrix factorization-based approach to reconstruct the missing traffic data [25]. Moreover, some researchers have expanded the two-dimensional matrix into high-dimensional tensors, such as the tensor decomposition models [26, 27]. However, the accuracy of these approaches mainly relies on the priori assumption of the data distribution, but the unknown data of the actual distribution may cause errors.

2.2. Deep Learning-Based Imputation. Recently, the booming of deep learning [28, 29] has inspired new ideas for data imputation. The autoencoder models [30] were developed to hierarchically train the full set of traffic data and extract the spatiotemporal features of the hidden layers to demon-

strate the effectiveness of data imputation [31]. Pathak et al. converted spatiotemporal trajectory data into images and used the powerful feature extraction by convolutional neural network (CNN) and then combined with the autoencoder model to impute the missing spatiotemporal trajectory data [32]. Generative adversarial network (GAN) provides a class of generative models for adversarial training, and it applies actual data/parallel data to generate the true data distribution, so that the imputation quality would be improved [33, 34]. By incorporating the reversibility of the generative imputer into GAN, Kazemi and Meidani proposed an iterative GAN architecture to evaluate the imputation of traffic missing data [35]. These state-of-the-art models work properly while dealing with the data correlation across different road segments. It adopts the powerful feature extraction capabilities of the deep neural networks to impute spatiotemporal data. However, most methods ignore the spatial dependencies in traffic network, and their imputation effect depends on a massive number of training data.

As some graph-based methods consider strong relations of data structures [36, 37], it is feasible of capturing global information to improve imputation performance. Chen and He proposed a heterogeneous graph embedding framework, which constructed a travel heterogeneous information network to find the best matched vehicles for the missing records [38]. By incorporating the spectral graph convolution operation, Cui et al. developed the graph Markov network to handle missing values for short-term traffic forecasting [39]. Graph representation learning is one crucial category of deep learning that has been widely used for traffic data imputation [40]. They viewed the observations and features as data nodes in a bipartite graph or constructed sample self-representation strategy and further required the neighbouring missing samples. Motifs are small connected components in a graph and are beneficial to understand the higher-order relations and global spatial graph principles [41, 42]. They introduced some imputation approaches based on motif discovery [43, 44], which are rarely adopted in transportation. The graph-based methods solve the imputation problems by capturing global spatial information from historical and neighbouring data and ignore the time series results to reflect the spatial dependencies.

Inspired by the above viewpoints, we propose a spatiotemporal imputation approach, which is based on motif-based graph aggregation. By adopting the motifs, the proposed approach benefits from capturing the higher-order spatial correlations in traffic networks. In addition, the

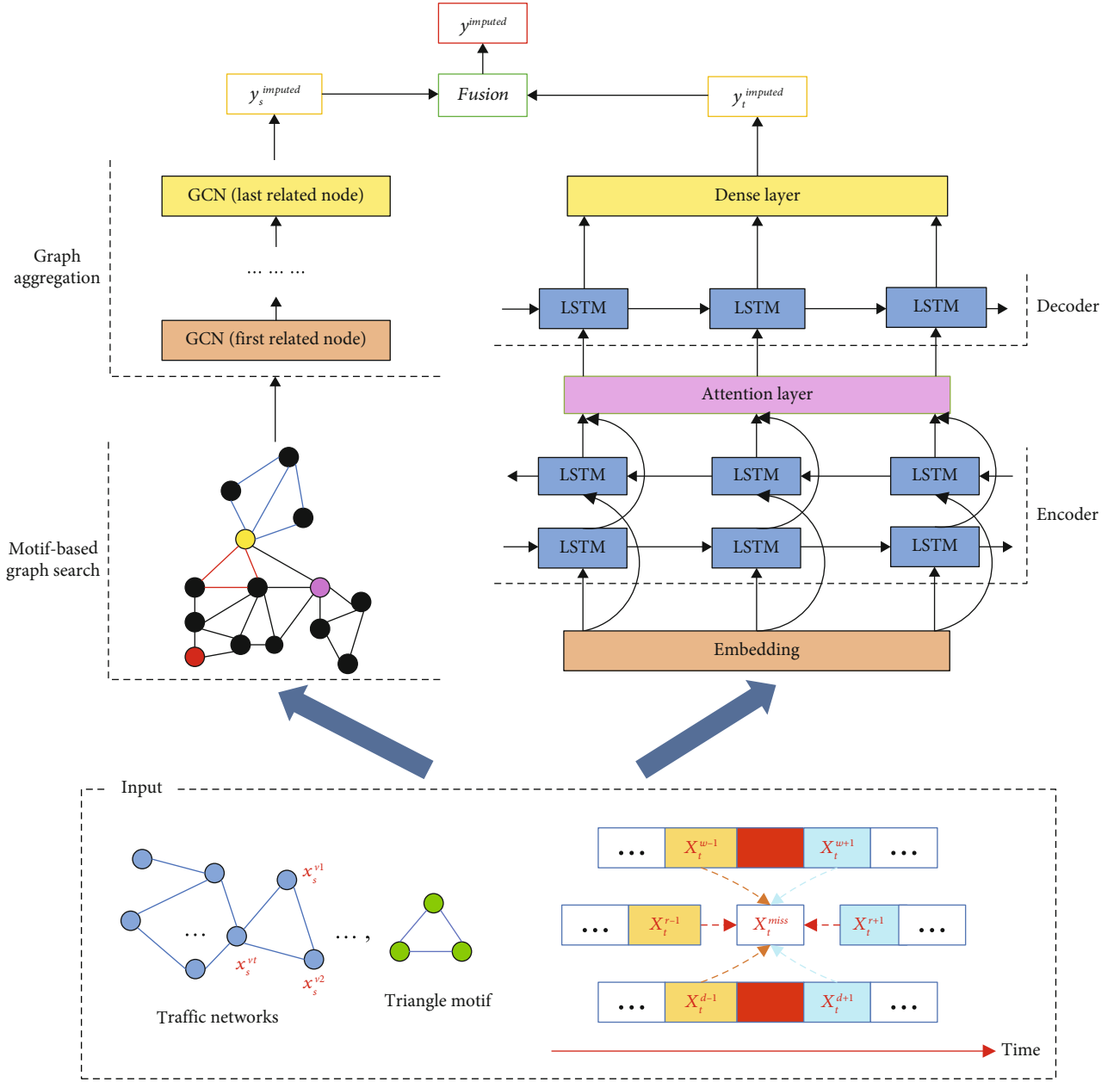


FIGURE 2: Framework of the MGIA. The bottom of figure indicates the data input, includes the higher-order spatial structure and time series data. The upper of figure indicates the specific method of the MGIA, the left module represents the spatial imputation of motif-based graph aggregation, the right module represents the multiple dimension imputation based on Bi-LSTM, and these two modules are integrated to obtain results.

proposed approach focuses on the imputation issue in the case of continuous data missing. At last, we list the limitations of these existing methods in Table 1.

3. Methodology

According to the analysis of related work, these existing methods are with the following limitations: (1) they ignore the spatial dependencies in the scene of large-scale areas and (2) rarely consider the imputation issue in the case of continuous data missing. The MGIA is aimed at improving the imputation accuracy from the perspectives of continuous

data missing and the higher-order spatial correlations in traffic networks. The framework of the MGIA is shown in Figure 2. It works in the following steps: (1) we adopt motifs to define the graph-based structure presented in traffic networks and search for all associated road segments that meet the motif gain condition of the missing data segment. On this basis, GCN is utilized to gradually aggregate the non-missing features of each associated segment, and the spatial aggregated value of the missing data segment is determined. (2) The multitime dimension imputation based on Bi-LSTM focuses on dealing with the problem of continuous data missing, which incorporates the recent, daily-periodic, and

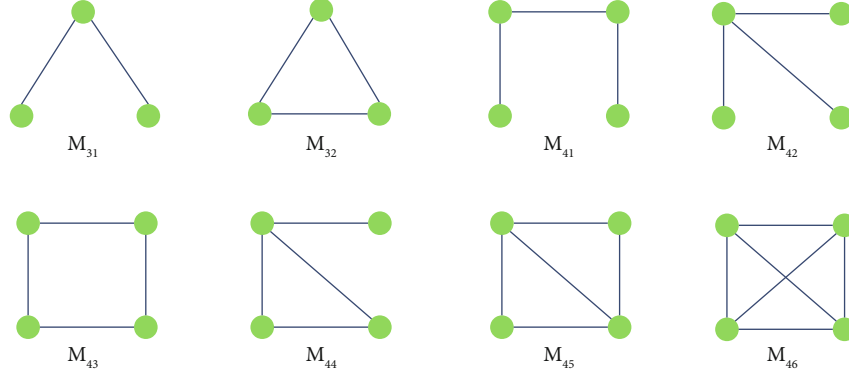


FIGURE 3: Examples of motifs. M_{31} and M_{32} denote triangles, and M_{41} , M_{42} , M_{43} , M_{44} , M_{45} , and M_{46} denote quadrangles.

weekly-periodic dependencies of the historical data. (3) The spatial aggregated values and the multitime dimension fusion values are integrated to obtain the imputation results.

3.1. The Spatial Imputation of Motif-Based Graph Aggregation. In order to capture the data correlation and global spatial characteristics between the road segments in traffic networks, the motif-based graph aggregation is employed to impute the missing data of road segments.

3.1.1. The Associated Node Search Based on Motif Discovery. Motifs are nonisomorphic connected graph structures that occur frequently in the network and the number of nodes is greater than or equal to 3, triangle and quadrangle motifs are shown in Figure 3. As motifs consider higher-order correlations of data structures, it is feasible for capturing global information to improve the performance of node feature aggregation [45]. Triangles are traffic network motifs that play important roles in the higher-order connectivity [46]. Thus, we select the triangle motif M_{32} as the research object in combination with traffic theory, in which nodes denote road segments, and edges denote the connection between two adjacent segments.

We design the method of motif-based search, which uses the motif gain to adjust the fitness function. If the motif gain no longer increases or no neighbouring node exists, then the search phase is stopped.

First, a road segment set $Q_v = \{v_1, v_2, \dots, v_n\}$ is defined in traffic networks. Assuming that the road segment where the missing data in Q_v is a node v_t , a target node set $Q_t = \{v_t\}$ is determined. Then, the motif-based local optimization algorithm is adopted to search all adjacent nodes of node v_t and form adjacent node set $Q_n = \{v_1^n, v_2^n, \dots\}$ and calculates the motif gain owing to adding node v_i^n in node set Q_n to node set Q_t . If the motif gain is greater than zero (i.e., there is a spatial correlation between the node v_t and its adjacent nodes), the node v_t will be added to the target node set Q_t and $Q_t = \{v_t, v_i^n\}$ is updated. Third, by adding the remaining adjacent nodes of Q_t , the motif gain is calculated in turn, and the nodes that meet the conditions are joined to the node set Q_t , $Q_t = \{v_t, v_1^n, v_2^n, \dots\}$ is updated at last. By analogy, the eligible nodes continue to join the target node set Q_t . When the motif gain calculation of all nodes in the node set $Q_v = \{v_1, v_2, \dots, v_n\}$ is completed, the associated node search for node

v_t ends. The associated node search for node v_t is shown in Figure 4.

The local motif rate addresses the issue of avoiding repeated counting motifs [47]. When an adjacent node v_i^n is joined, the local motif rate is calculated as follows:

$$R_M(Q \cup v_i^n) = \frac{N_{in}(Q) + E_{in}(v_i^n)}{(N_{in}(Q) + N_{out}(Q) + E_{out}(v_i^n))^\alpha}, \quad (1)$$

$$R_M(Q) = \frac{N_{in}(Q)}{(N_{in}(Q) + N_{out}(Q))^\alpha}, \quad (2)$$

where Q is the current node set, v_i^n is the joined i^{th} adjacent node, $R_M(\cdot)$ is the local motif rate, $N_{in}(\cdot)$ is the number of local motif, $N_{out}(\cdot)$ is the number of motifs between the current node set and the external node set, $E_{in}(\cdot)$ is the number of motifs between the current node set and the new nodes, and $E_{out}(\cdot)$ is the number of motifs between the external node set and the new nodes.

The motif gain of the node v_t is calculated as follows:

$$R_M^{v_i^n}(Q \cup v_i^n) = R_M(Q \cup v_i^n) - R_M(Q), \quad (3)$$

where $R_M^{v_i^n}(\cdot)$ is the gain of the local motif rate when the node v_i^n is joined.

The gain $R_M^{v_i^n} > 0$ indicates that the joined adjacent node v_i^n is spatially related to the missing data node v_t . As shown in Figure 5, we provide two examples of calculating the motif gain $R_M^{v_i^n}$. $N_{in}(Q)$ and $N_{out}(Q)$ are equal to 4 and 3, respectively, and $R_M(Q)$ is $4/(4+3)^\alpha$. In Figure 5(a), $E_{in}(v_1)$ and $E_{out}(v_1)$ are equal to 1 and 2, respectively, and $R_M(Q \cup v_1)$ is $(4+1)/(4+3+2)^\alpha$ from Equation (3), the default control parameter α is 1, and $R_M^{v_1} < 0$; thus, the node v_1 cannot be joined to the current node set Q . In Figure 5(b), $E_{in}(v_2)$ and $E_{out}(v_2)$ are equal to 1, and $R_M(Q \cup v_2)$ is $(4+1)/(4+3+1)^\alpha$, $R_M^{v_2} > 0$; this indicates that there is a spatial correlation between the missing data node v_t and the node v_2 .

3.1.2. The Spatial Aggregated Method Based on GCN. All nodes associated with the missing data node v_t by Equation (3) are determined, and the associated node data is aggregated to obtain the estimate values of the missing data node

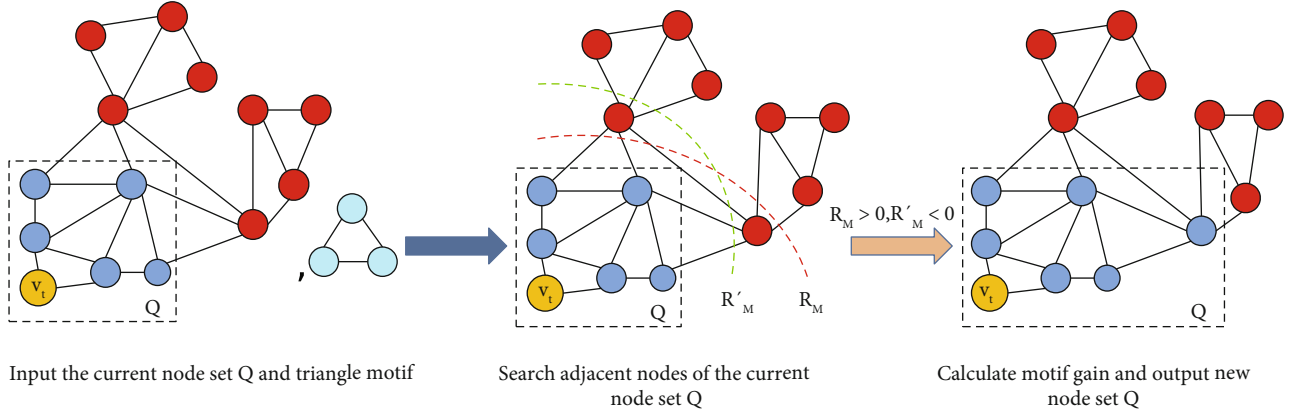
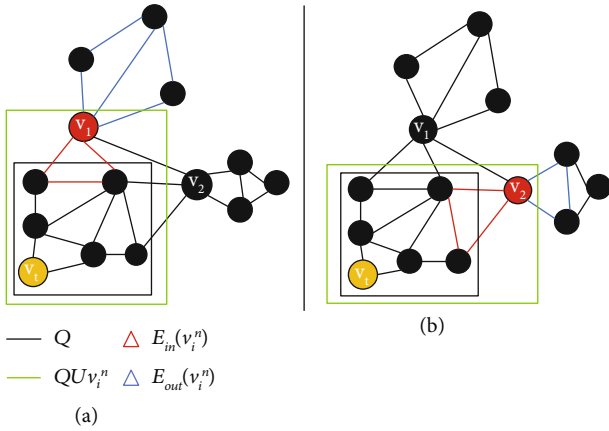


FIGURE 4: The process of motif-based search.

FIGURE 5: Examples of calculating local motif rates. (a) Calculating $R_M(Q \cup v_1)$. (b) Calculating $R_M(Q \cup v_2)$.

v_t by GCN. A graph $G = (V, E, W)$, with n nodes to describe a road network, where nodes $v_i \in V$ denote road segments, edges $(v_i, v_j) \in E$ denote the directed connection from node v_i to node v_j and $W \in R^{n \times n}$ denotes the weighted adjacency matrix. The graph G is represented by its corresponding Laplacian matrix. The properties of the graph structure can be obtained by analyzing Laplacian matrix and its eigenvalues.

$$L = I_N - D^{-1/2}AD^{-1/2}, \quad (4)$$

$$g_\theta * x = Ug_\theta U^T x, \quad (5)$$

where L is the normalized form of Laplacian matrix and D , A , and I_N are the degree matrix, adjacent matrix, and unit matrix, respectively. g_θ and U are the eigenvector function and matrix of L , respectively, and x is the eigenvalue of input node.

According to Equations (4) and (5), the eigenvalue decomposition of L is represented as follows:

$$L = U\Lambda U^T, \quad (6)$$

where Λ is the diagonal matrix composed by eigenvalues of L and $\Lambda = \text{diag}([\lambda_0, \lambda_1, \dots, \lambda_{N-1}]) \in R^{N \times N}$.

In order to reduce the time complexity when the scale of the graph is large, the Chebyshev polynomials are adopted to approximate the solution:

$$g_\theta * x \approx \sum_{k=0}^K \theta_k T_k(\tilde{L})x, \quad (7)$$

$$T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x), \quad (8)$$

where $T_k(x)$ and θ_k are Chebyshev polynomials and coefficients, respectively; $\tilde{L} = (2/\lambda_{\max})L - I_N$; λ_{\max} represents the maximum eigenvalue of L ; $T_0(x) = 1$; and $T_1(x) = x$.

It can be seen from Equations (7) and (8), the approximate solution with Chebyshev polynomials is equivalent to using a convolution kernel to extract the eigenvalues of neighboring nodes with the node as the center of each node in the graph. In order to simplify the calculation, limit k to 1, scale the eigenvalue of L to make $\lambda_{\max} = 2$, and Equation (7) is expressed as follows:

$$g_\theta * x \approx \theta_0 x + \theta_1 (L - I_N)x. \quad (9)$$

According to Equation (4), let $\theta_0 = -\theta_1 = \theta$ at the same time, and Equation (9) is transformed as follows:

$$g_\theta * x \approx \theta (I_N + D^{-1/2}AD^{-1/2})x, \quad (10)$$

In order to avoid numerical and gradient instability problems, let $\tilde{A} = I_N + A$, Equation (10) is transformed as follows:

$$g_\theta * x \approx \theta (D^{-1/2}\tilde{A}D^{-1/2})x. \quad (11)$$

All associated nodes that meet the motif gain conditions are extracted according to Equation (11), and final aggregated value of the missing data node v_t is expressed as follows:

```

Input: the node set  $Q_v = \{v_1, v_2, \dots, v_n\}$ , target node  $v_t$ , the eigenvalue of the first associated node  $y_s^{(0)}$ .
Output: imputed value  $y_s^{\text{imputed}}$ .
Part I: Motif-based search.
1: Initialize: current node set  $Q = \{v_t\}$ , motif gain  $R_M^{v_t} = 0$ 
2: repeat
3: while  $R_M^{v_t} > 0$  do
4:   update  $Q = \{v_t, v_i^n\}$ 
5:   set  $R_M(Q)$  according to Equation (2)
6:   obtain  $R_M(Q \cup v_{i+1}^n)$  according to Equation (1)
7:   obtain  $R_M^{v_{i+1}^n}(Q \cup v_{i+1}^n)$  according to Equation (3)
8:    $i \leftarrow i + 1$ 
9: end while
10: until no neighbor node exists
11: update target node set  $Q_t = Q$ 
Part II: Aggregation
12: Initialize:  $y_s^{\text{imputed}} = y_s^{(0)}$ 
13: for  $v_i^n \in Q_t$  do
14:   update  $y_s^{\text{imputed}}$  according to Equations (4)–(12)
15: end for
16: return  $y_s^{\text{imputed}}$ 

```

ALGORITHM 1: Motif-based graph aggregation.

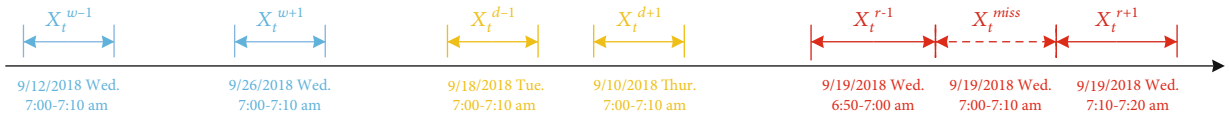


FIGURE 6: The construction of adjacent time series data in three dimensions.

$$y_s^{(l)} = \sigma \left(D^{-1/2} \tilde{A} D^{-1/2} y_s^{(l-1)} W^{(l-1)} \right), \quad (12)$$

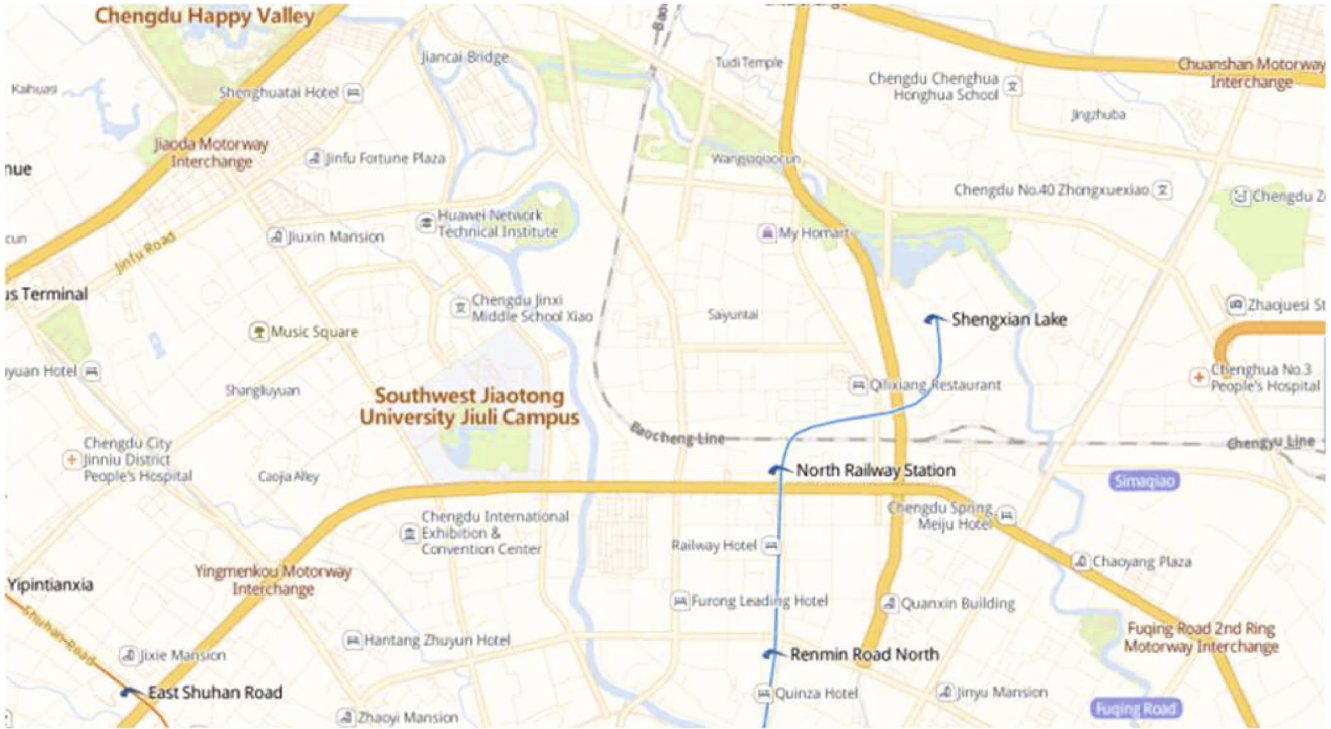
where $y_s^{(l)}$ is the final aggregated value of the missing data node v_t , i.e., the spatial aggregated value y_s^{imputed} in Figure 2. l , W , and σ are the number of associated nodes, the parameters to be trained, and activation function, respectively. The initial value $y_s^{(0)}$ is the eigenvalue of the first associated node that meets the motif gain condition.

3.1.3. The Computational Complexity. We discuss the time complexity of the proposed approach. For the process of motif-based node search, a node v_t where the missing data needs to search other $n - 1$ nodes, and the nodes in the target node set Q_t also search the each node that do not exist in Q_t , so the computational complexity is $O(n^2)$. For the process of the GCN-based aggregation, optimized by the Chebyshev polynomials, the time complexity is reduced to $O(k|E|)$. In order to simplify the calculation, we limit k to 1, so the time complexity of the GCN-based aggregation approximates to $O(n)$. It is a sequential to these two processes, so the total value of the two processes is $O(n^2 + n)$, i.e., the overall computational complexities $O(n^2)$.

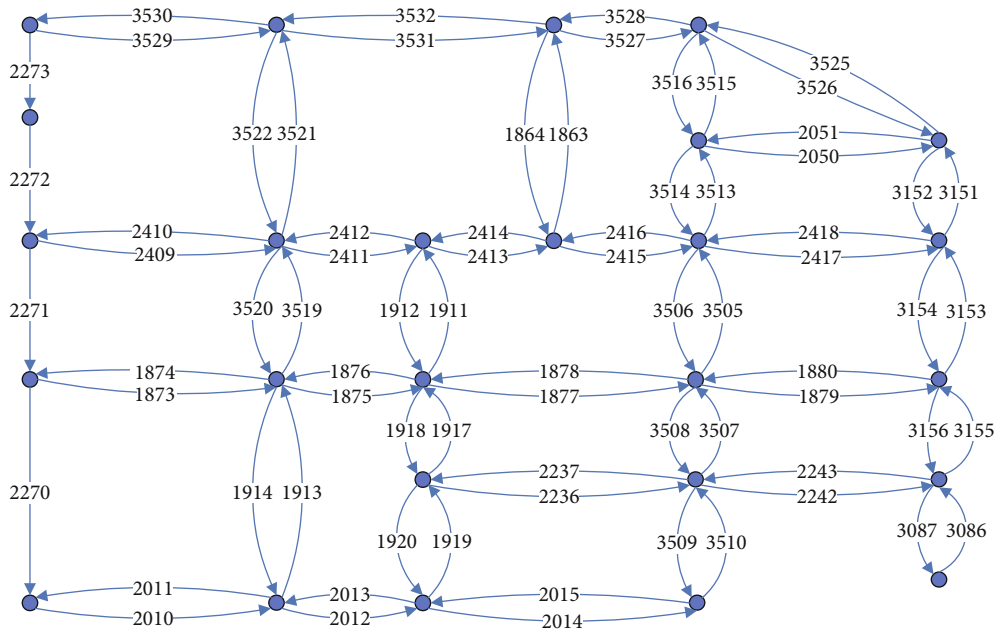
The process of motif-based graph aggregation is depicted in Algorithm 1.

3.2. The Multiple Dimension Imputation Based on Bi-LSTM. Existing temporal imputation methods obtain good estimates of missing data in the case of random data missing. When there are a large number of continuous missing values at some consecutive timestamps, the imputation performance will degrade [15]. A more challenging task is to recover the continuous missing data.

Processing long-term time-series data is an essential task since there is a large number of continuous missing data. Deep learning, which trains classifiers directly from input data by complex feature representations, may generate high performing results in the dynamic and challenging context [48]. LSTM and gated recurrent unit (GRU) [49], which are the key methods of deep learning, have been employed for time-series applications with temporal dependencies. LSTM has three gates, in which, input gate and forget gate are used to control the update of memory cell, and output gate passes the output information. LSTM has recently been employed for missing data imputation, such as in Ref. [10, 13, 15]. GRU is the variant of LSTM that only comprises of update gate and reset gate and utilizes the reset gate to control the information at the previous point. GRU's training of previous point could restrict improvement owing to the case of continuous data missing. Besides, in the case of a large amount of dataset, the performance of LSTM is better than that of GRU, and it has been verified by [50]. Therefore, in order to efficiently capture the temporal dependencies of



(a)



(b)

FIGURE 7: Study area description. (a) Geographical area in Chengdu, China. (b) The road network in this work.

traffic time series data, the LSTM is employed in our proposed approach.

It is known that the missing data points are closely related to the adjacent points from two opposite temporal directions. Bi-LSTM is capable of training in both forward and backward directions [51]. Meanwhile, an increase of the window based on different time granularity can provide benefit in prediction performance, by allowing the capitali-

zation of temporal dependencies in the time series data [29]. Thus, we adopt the time-series imputation based on Bi-LSTM, which incorporates the recent, daily-periodic, and weekly-periodic dependencies of the historical data. That is to say, among the adjacent time-series data of current missing values in the three dimensions, there exist normal data at least one dimension, so as to ensure that there are enough historical data to complete the temporal imputation.

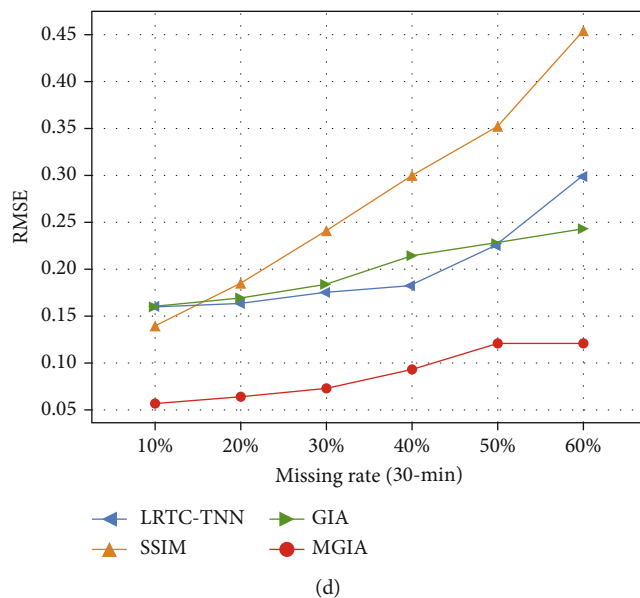
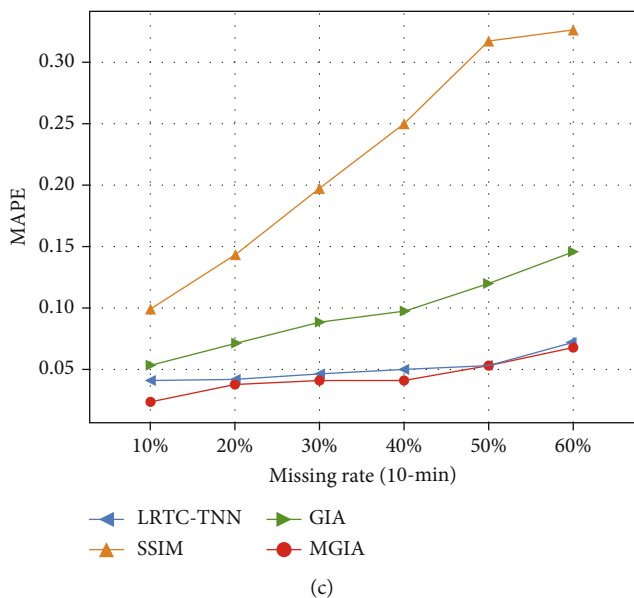
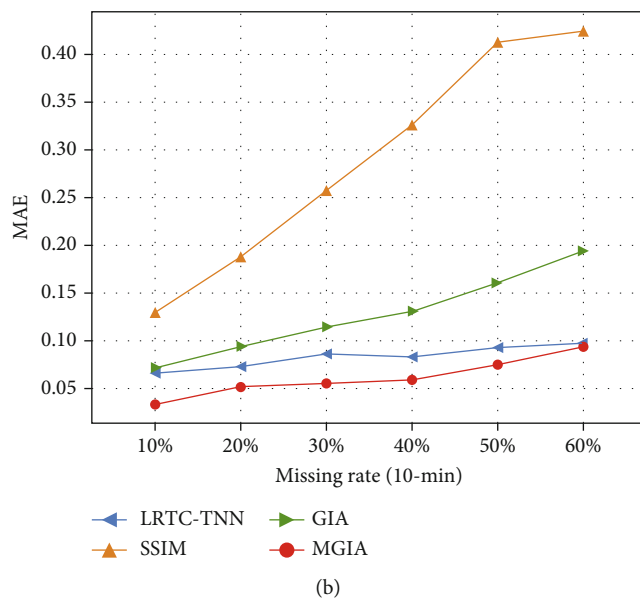
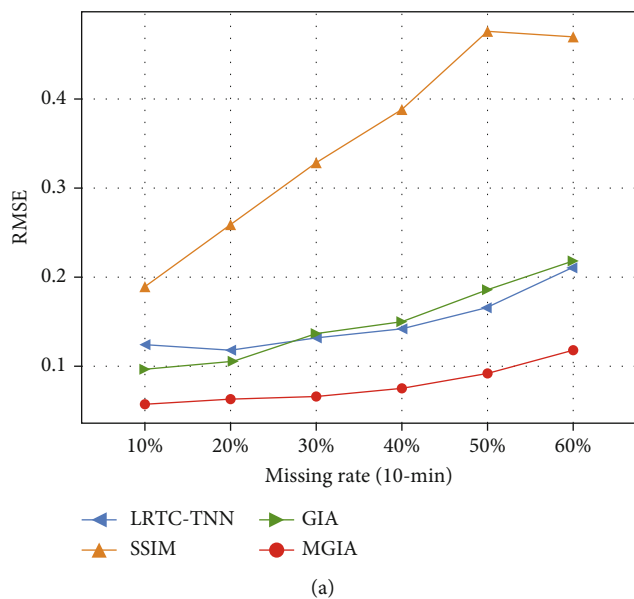


FIGURE 8: Continued.

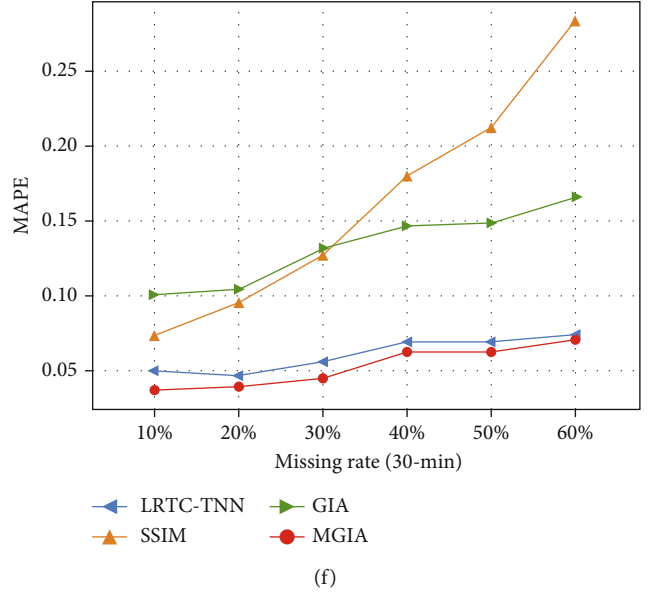
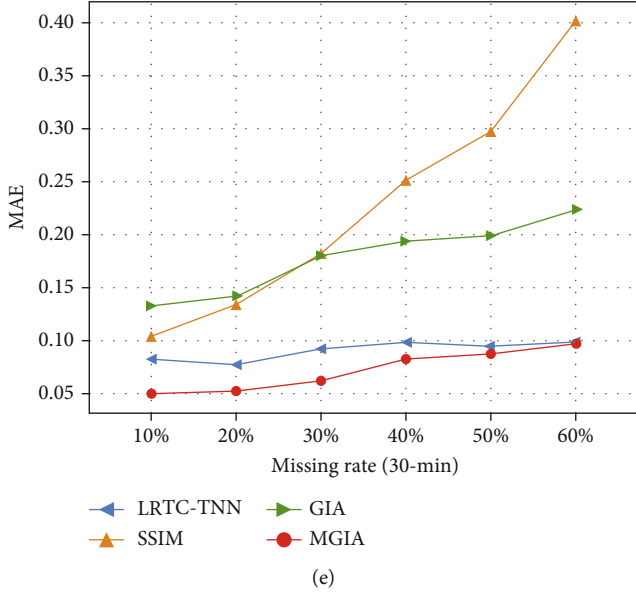


FIGURE 8: Metrics in the case of random data missing. (a–c) The results of RMSE, MAE, and MAPE with 10 min time interval, respectively. (d–f) The results of RMSE, MAE, and MAPE with 30 min time interval, respectively. The abscissa represents the missing rate, and the ordinate is the corresponding result.

Assuming that the road segment has missing data x_t^{miss} in time period t , the adjacent data of recent, daily, and weekly periods are defined as $[x_t^{r-n}, \dots, x_t^{r-1}, x_t^{r+1}, \dots, x_t^{r+n}]$, $[x_t^{d-n}, \dots, x_t^{d-1}, x_t^{d+1}, \dots, x_t^{d+n}]$, and $[x_t^{w-n}, \dots, x_t^{w-1}, x_t^{w+1}, \dots, x_t^{w+n}]$, respectively. For instance, when x_t^{miss} represents the data from 7 a.m. to 7:10 a.m. on 19 September 2018, some adjacent time series data in three dimensions are shown in Figure 6.

3.2.1. Bi-LSTM Encoder. The Bi-LSTM extends the generic LSTM. In case of recent, daily, and weekly periods, it utilizes the previous and future points by processing the missing data x_t^{miss} from both forward and backward directions with two separate LSTMs. The one-way LSTM of recent, daily, and weekly periods include forget, input, and output gates.

In each time dimension, the hidden vector h_t in time period t is updated as follows:

$$f_t = \sigma(W_f h_{t-1} + U_f x_t^{\text{adj}} + b_f), \quad (13)$$

$$i_t = \sigma(W_i h_{t-1} + U_i x_t^{\text{adj}} + b_i), \quad (14)$$

$$a_t = \tan h(W_a h_{t-1} + U_a x_t^{\text{adj}} + b_a), \quad (15)$$

$$C_t = C_{t-1} \otimes f_t + i_t \otimes a_t, \quad (16)$$

$$o_t = \sigma(W_o h_{t-1} + U_o x_t^{\text{adj}} + b_o), \quad (17)$$

$$h_t = o_t \otimes \tan h(C_t), \quad (18)$$

where f_t and o_t refer to forget and output gates, respectively, i_t and a_t are input gates, x_t^{adj} is the adjacent historical data in

time period t , and C is the cell vector. σ , $\tan h$, and \otimes represent the sigmoid function, $\tan h$ activation function, and element-wise multiplication, respectively.

The previous and future hidden vectors generated by Equations (13)–(18) are updated in the time period t :

$$h_t = h_{t,p} + (1 - \theta)h_{t,f}, \quad (19)$$

where $h_{t,p}$ and $h_{t,f}$ represent the previous and future hidden vectors, respectively, and θ is the impact weight.

3.2.2. LSTM Decoder with Attention. In this part, we select the LSTM decoder with the attention mechanism. The encoded hidden vectors of adjacent periods are fused into an attention vector:

$$c_t = \sum_{i=0}^n \alpha_{t-i} h_{t-i} + \sum_{i=1}^n \alpha_{t+i} h_{t+i}, \quad (20)$$

$$\alpha_t = \text{softmax}(f(s_{t-1}, h_t)),$$

where c_t is the attention context vector in time period t , $h(\cdot)$ is the updated hidden vector in encoding stage, s is the decoded hidden vector, f is the forward LSTM trained by decoder components, and α and softmax are weight coefficient and activation function, respectively.

The dense layer with a linear activation function is added on top of the decoder layer to generate predictions in Figure 2 and outputs the decoded values with the backward LSTM:

$$y_t = \text{Linear}(W[s_t; c_t] + b), \quad (21)$$

$$s_t = \text{LSTM}(y_{t-1}, s_{t-1}, c_t),$$

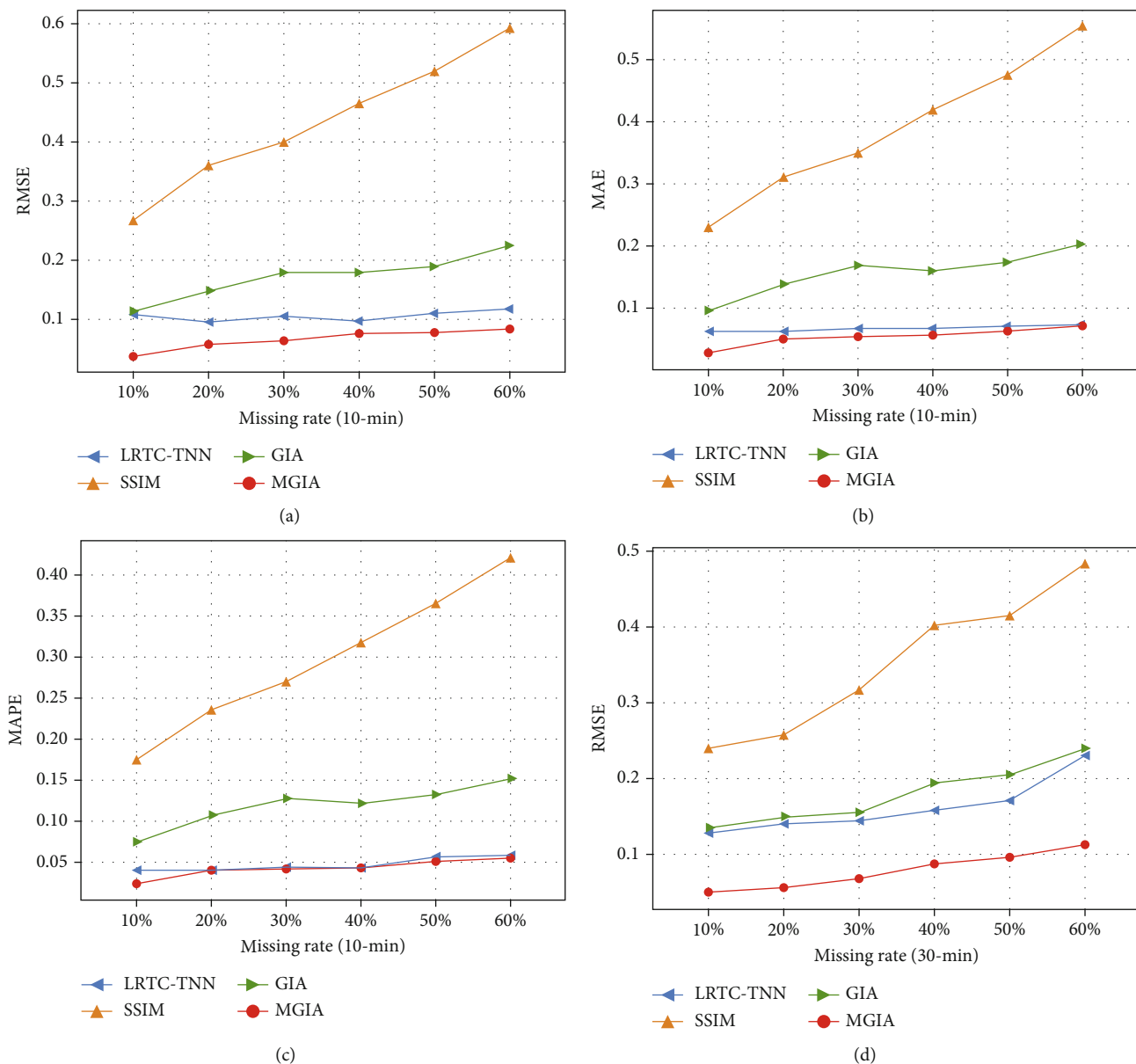


FIGURE 9: Continued.

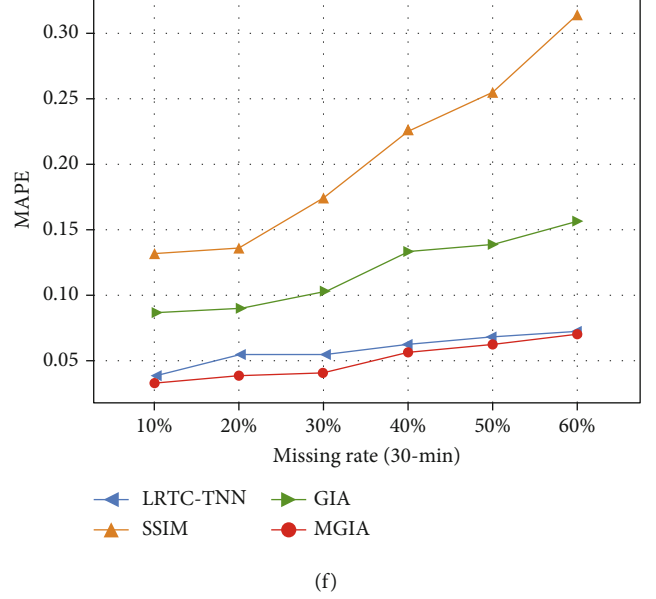
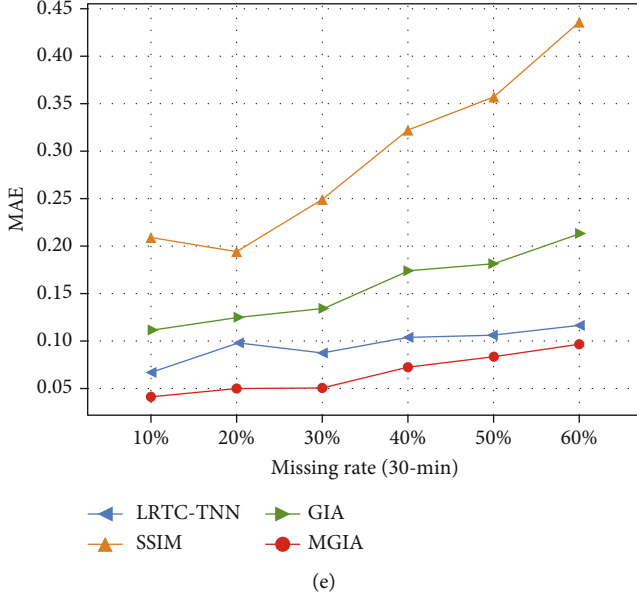


FIGURE 9: Metrics in the case of continuous data missing. (a–c) The results of RMSE, MAE, and MAPE with 10 min time interval, respectively. (d–f) The results of RMSE, MAE, and MAPE with 30 min time interval, respectively. The abscissa represents the missing rate, and the ordinate is the corresponding result.

where y is the output value decoded by LSTM, $[s_t; c_t]$ is a concatenation of the decoder hidden vector s_t and the attention context vector c_t , Linear is linear activation function, and W and b are linear parameters mapped to decoder hidden states.

3.2.3. Multicomponent Time Dimension Fusion. The input vectors $[x_t^{w-1}, x_t^{d-1}, x_t^{r-1}, x_t^{r+1}, x_t^{d+1}, x_t^{w+1}]$ are trained by the above encoder-decoder processes, and the estimated value merged by the adjacent data of the missing data in the three time dimensions of recent, daily and weekly periods, i.e., the fusion-value y_t^{imputed} are obtained.

3.3. The Spatiotemporal Fusion of Missing Data. The spatial aggregated value y_s^{imputed} and the multitime dimension fusion-value y_t^{imputed} in time period t are combined as follows:

$$y^{\text{imputed}}(t) = a(t)y_s^{\text{imputed}}(t) + (1 - a(t))y_t^{\text{imputed}}(t), \quad (22)$$

where $y^{\text{imputed}}(t)$ is the final result of spatiotemporal imputation in time period t , and the initial value of weight $a(t) = 1/2$.

In order to prevent the poor effect caused by excessive weight fluctuation, an inertia factor β is introduced. The size of β reflects the size of the weight inertia. The β is inversely proportional to the volatility of weight a .

Supposing that the error of the spatial imputation in time period t is $e_{1,t}$, and the error of the multitime dimension imputation in time period t is $e_{2,t}$, $a(t)$ is updated as follows:

$$a(t) = \beta \times a(t-1) + (1 - \beta) \frac{|e_{2,t}|}{|e_{1,t}| + |e_{2,t}|}. \quad (23)$$

That is to say, when the imputation error is relatively large, in the next time period, the imputation influence will be reduced and vice versa. The value of β can be obtained through simulation experiments to obtain a better value.

4. Results and Discussion

In this section, we introduce the dataset, baselines, and evaluation metrics, and then verify the superiority of the MGIA in the case of random and continuous data missing.

The scenario of random data missing is that missing data appears randomly in the traffic dataset owing to the temporary failure (e.g., network or power outage issues). Continuous data missing refers to that some data is missing continuously due to the long-term failure of traffic data collectors, in which the values are missed at some consecutive timestamps or multiple intersections in a road network.

4.1. Data Preparation. We validate our approach with the traffic index dataset, which was provided by the Chengdu branch of Didi Chuxing, China. In Figure 7, we select $30.66^\circ \text{ N} \sim 30.73^\circ \text{ N}$, $104.02^\circ \text{ E} \sim 104.10^\circ \text{ E}$ as the geographic area and 74 road segments make up the road network in this work. The time span of these dataset is from September to October 2018. All the programs are developed based on Python 3.7 and TensorFlow 1.13.1.

The dataset was filtered under the normal traffic flow hours (6 am–8:50 pm), and the time intervals are set to 10 min and 30 min, respectively. The traffic index is a

TABLE 2: Analysis of the imputation with MGIA (10 min time interval). The acronyms “r-missing” and “c-missing” represent the patterns of random data missing and continuous data missing, respectively.

Pattern	Metrics	Missing rate					
		0.1	0.2	0.3	0.4	0.5	0.6
r-missing	RMSE	0.0562	0.0622	0.0659	0.0746	0.0914	0.1176
c-missing	RMSE	0.0371	0.057	0.0631	0.0733	0.0759	0.0843
r-missing	MAE	0.0325	0.0501	0.0547	0.0581	0.0742	0.0938
c-missing	MAE	0.0292	0.0508	0.0535	0.0563	0.0653	0.0726
r-missing	MAPE	2.51%	3.81%	4.17%	4.27%	5.45%	6.95%
c-missing	MAPE	2.26%	3.94%	4.02%	4.2%	4.97%	5.45%

TABLE 3: Analysis of the imputation with MGIA (30 min time interval). The acronyms “r-missing” and “c-missing” represent the patterns of random data missing and continuous data missing, respectively.

Pattern	Metrics	Missing rate					
		0.1	0.2	0.3	0.4	0.5	0.6
r-missing	RMSE	0.0548	0.0625	0.072	0.0939	0.121	0.1231
c-missing	RMSE	0.048	0.0534	0.0657	0.0854	0.0946	0.1106
r-missing	MAE	0.0492	0.0522	0.062	0.0821	0.0862	0.0974
c-missing	MAE	0.0412	0.051	0.052	0.0731	0.082	0.0967
r-missing	MAPE	3.71%	3.9%	4.47%	6.24%	6.37%	7.11%
c-missing	MAPE	3.24%	3.74%	3.96%	5.61%	6.23%	6.98%

standardized quantitative indicator to measure traffic congestion. It is expressed as follows:

$$C_i^t = \frac{v_i^{\max} - v_i^t}{v_i^{\max} - v_i^{\min}}, \quad (24)$$

where C_i^t is the traffic index of road segment i during time-slot t , v_i^t is the average speed, and v_i^{\max} and v_i^{\min} are the maximum speed and minimum speed corresponding to road segment i in the historical data, respectively.

In the tests, we manually remove a certain amount of processed dataset and then compute these data with the proposed approach and state-of-the-art methods. We assume that there are patterns of random and continuous data missing in the dataset and get the accuracy of these methods by comparing the imputed results with the ground truth data. The range of data missing rate in these two patterns is set between 10% and 60% (in units of 10%).

4.2. Experimental Settings

4.2.1. *Baselines.* The comparison between the MGIA and state-of-the-art methods for traffic data imputation is conducted as follows:

- (i) LRTC-TNN [14] is a low-rank tensor completion framework with truncated nuclear norm, which extracts spatiotemporal features from traffic data
- (ii) SSIM [15] is a sequence-to-sequence imputation model, which is designed to impute missing data by utilizing the LSTM from both the past and future time indexes

(iii) MGIA: our proposed approach, which incorporates the motif-based graph aggregation method with the multitime dimension fusion method based on Bi-LSTM, to impute missing data

(iv) GIA is a comparison for MGIA, which incorporates the graph aggregation method with the multitime dimension fusion method based on Bi-LSTM but does not include motif-based application

4.2.2. *Evaluation Metrics.* To evaluate the performance of the MGIA, we employ three widely used performance metrics: root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE).

$$\begin{aligned} \text{RMSE} &= \sqrt{\frac{1}{n} \sum_i^n (x_i^r - y_i)^2}, \\ \text{MAE} &= \frac{1}{n} \sum_i^n |x_i^r - y_i|, \\ \text{MAPE} &= \frac{1}{n} \sum_i^n \frac{|x_i^r - y_i|}{x_i^r}, \end{aligned} \quad (25)$$

where n is the total number of the missing data, y_i is the i^{th} imputed data, and x_i^r is the corresponding actual data of y_i .

4.3. *Comparison and Result Analysis.* As illustrated in Figures 8 and 9, the average errors with different missing rates in the case of random and continuous data missing is presented, respectively. That shows the MGIA shows superior performance gains over the baselines w.r.t. all the three metrics.

All metrics of MGIA are lower than those based on the other three approaches. LRTC-TNN and SSIM are the recent approaches, and the difference between all three metrics of MGIA and those based on other two algorithms are larger, which means that the experiments based on MGIA achieved good performance. Meanwhile, MGIA significantly outperforms the compared approach GIA. It means that the application of motif in spatial imputation is feasible and effective.

As shown in Figures 8 and 9, the overall trends of the four approaches are almost similar, i.e., all metrics of these approaches increase with the increase in the missing rate. In Figure 8, when the missing rate is less than 40%, the error growth of these approaches is steady except for SSIM. Once the missing rate exceeds 40%, the error increase is larger than the respective miss rate (the missing rate is less than 40%). In Figure 9, the error growth of these approaches has remained stable, which crosses various missing rates from 10% to 60%. The reason could be that, as for random data missing, there will be the situation of continuous data missing when the missing rate exceeds 40%; thus, this is different from the current imputed pattern, and the error growth increases significantly.

The LRTC-TNN achieved a better performance than the SSIM. This is because LRTC-TNN makes use of low-rank tensor decomposition and will not be affected by consecutive missing data. On the contrary, SSIM is susceptible to consecutive missing data. If the data is missing at the forward period or the next period, the error will be very different. GIA improves SSIM from one-dimensional to three-dimensional horizons and employs graph aggregation to perform spatiotemporal fusion of data imputation, so it has improved performance compared to SSIM. MGIA adds motif-based imputation on the basis of GIA and balances the effectiveness of higher-order spatial correlations and periodicity in merging the imputed values.

Due to the widening of the time interval, i.e., from 10 min to 30 min, the imputation performance becomes worse in Figures 8 and 9, and MGIA still performs better than the other three approaches in all the three metrics.

To represent the advantage of continuous data missing in MGIA, we compare the metrics with the pattern of random data missing in Tables 2 and 3. Regardless of whether it is a 10 min interval or a 30 min interval, the pattern of continuous data missing has better performance than the pattern of random data missing. This is because MGIA adopts the time series method based on Bi-LSTM, which widens the horizon from one-dimension to three-dimension, and pays more attention to the time continuity.

Overall, MGIA outperforms the other baselines due to the spatiotemporal characteristics and higher-order spatial correlations. Moreover, the imputation performance of continuous data missing is better than random data missing.

5. Conclusion

In this paper, a novel spatiotemporal imputation approach for traffic data (MGIA), which utilizes motif-based graph aggregation, is proposed. To sum up, this approach

addresses the issues of (1) traffic spatial imputation for large-scale areas and (2) poor imputation performance especially when in the case of continuous data missing. Based on MGIA, we capture the higher-order spatial correlations in traffic networks and solve the problem of spatiotemporal data imputation. Experiments are performed on a real-world traffic dataset in Chengdu, China. The experimental results showed that the MGIA outperformed all other methods in the case of random and continuous data missing and achieved strong stability crossing various missing rates range from 10% to 60%.

In the future, we will further evaluate the MGIA with regard to other factors (such as weather and event). Besides, we plan to incorporate adversarial learning into the proposed approach to improve the imputation accuracy, in the case of continuous data missing with longer time interval. On this basis, we improve the accuracy at the various missing rate range from 60% to 80%. Additionally, we plan to apply the proposed approach to other tasks in ITS, such as traffic prediction and causal discovery of the congestion propagation patterns.

Data Availability

All datasets in this study can be downloaded from <https://outreach.didichuxing.com>.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Nos. 62073295 and 62072409), Zhejiang Provincial Natural Science Foundation (LR21F020003), and Fundamental Research Funds for the Provincial Universities of Zhejiang (RF-B2020001).

References

- [1] H. P. Li, M. Li, X. Lin, F. He, and Y. H. Wang, "A spatiotemporal approach for traffic data imputation with complicated missing patterns," *Transportation Research Part C: Emerging Technologies*, vol. 119, article 102730, 2020.
- [2] D. F. Zhu, G. J. Shen, D. Y. Liu, J. J. Chen, and Y. J. Zhang, "FCG-ASpredictor: an approach for the prediction of average speed of road segments with floating car GPS data," *Sensors*, vol. 19, no. 22, p. 4967, 2019.
- [3] X. J. Kong, K. L. Wang, M. L. Hou et al., "A federated learning-based license plate recognition scheme for 5G-enabled internet of vehicles," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 12, pp. 8523–8530, 2021.
- [4] X. J. Kong, S. Q. Tong, H. R. Gao et al., "Mobile edge cooperation optimization for wearable internet of things: a network representation-based framework," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 7, pp. 5050–5058, 2021.
- [5] W. L. Ding and Z. F. Zhao, "DS-harmonizer: a harmonization service on spatiotemporal data stream in edge computing

- environment,” *Wireless Communications & Mobile Computing*, vol. 2018, article 9354273, 2018.
- [6] C. F. Gong and Y. Y. Zhang, “Urban traffic data imputation with detrending and tensor decomposition,” *IEEE Access*, vol. 8, pp. 11124–11137, 2020.
 - [7] D. G. Zhang, T. Zhang, Y. Dong, X. H. Liu, Y. Y. Cui, and D. X. Zhao, “Novel optimized link state routing protocol based on quantum genetic strategy for mobile learning,” *Journal of Network and Computer Applications*, vol. 122, pp. 37–49, 2018.
 - [8] A. Chaudhry, W. Li, A. Basri, and F. Patenaude, “A method for improving imputation and prediction accuracy of highly seasonal univariate data with large periods of missingness,” *Wireless Communications & Mobile Computing*, vol. 2019, article 4039758, 2019.
 - [9] X. Y. Jia, X. Y. Dong, M. Chen, and X. H. Yu, “Missing data imputation for traffic congestion data based on joint matrix factorization,” *Knowledge-Based Systems*, vol. 225, article 107114, 2021.
 - [10] L. C. Li, J. Zhang, Y. G. Wang, and B. Ran, “Missing value imputation for traffic-related time series data based on a multi-view learning method,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 8, pp. 2933–2943, 2019.
 - [11] Y. F. Zhuang, R. M. Ke, and Y. H. Wang, “Innovative method for traffic data imputation based on convolutional neural network,” *IET Intelligent Transportation Systems*, vol. 13, no. 4, pp. 605–613, 2019.
 - [12] B. Bae, H. Kim, H. Lim, Y. Liu, L. D. Han, and P. B. Freeze, “Missing data imputation for traffic flow speed using spatio-temporal cokriging,” *Transportation Research Part C: Emerging Technologies*, vol. 88, pp. 124–139, 2018.
 - [13] D. Li, L. H. Li, X. L. Li, Z. Ke, and Q. Hu, “Smoothed LSTM-AE: a spatio-temporal deep model for multiple time-series missing imputation,” *Neurocomputing*, vol. 411, pp. 351–363, 2020.
 - [14] X. Y. Chen, J. M. Yang, and L. J. Sun, “A nonconvex low-rank tensor completion model for spatiotemporal traffic data imputation,” *Transportation Research Part C: Emerging Technologies*, vol. 117, article 102673, 2020.
 - [15] Y. F. Zhang, P. J. Thorburn, W. Xiang, and P. Fitch, “SSIM-a deep learning approach for recovering missing time series sensor data,” *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 6618–6628, 2019.
 - [16] B. Ghosh, B. Basu, and M. O’Mahony, “Bayesian time-series model for short-term traffic flow forecasting,” *Journal of Transportation Engineering*, vol. 133, no. 3, pp. 180–189, 2007.
 - [17] M. Castro-Neto, Y. S. Jeong, M. K. Jeong, and L. D. Han, “Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions,” *Expert Systems with Applications*, vol. 36, no. 3, pp. 6164–6173, 2009.
 - [18] J. Ahn, E. Ko, and E. Y. Kim, “Highway traffic flow prediction using support vector regression and Bayesian classifier,” in *Proceedings of IEEE International Conference on Big Data and Smart Computing*, pp. 239–244, New York, USA, 2016.
 - [19] J. J. Tang, G. H. Zhang, Y. H. Wang, H. Wang, and F. Liu, “A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation,” *Transportation Research Part C: Emerging Technologies*, vol. 51, pp. 29–40, 2015.
 - [20] Q. Shang, Z. S. Yang, S. Gao, and D. R. Tan, “An imputation method for missing traffic data based on FCM optimized by PSO-SVR,” *Journal of Advanced Transportation*, vol. 2018, Article ID 2935248, 2018.
 - [21] W. Yin, P. Murray-Tuite, and H. Rakha, “Imputing erroneous data of single-station loop detectors for nonincident conditions: comparison between temporal and spatial methods,” *Journal of Intelligent Transportation Systems*, vol. 16, no. 3, pp. 159–176, 2012.
 - [22] G. Chang, Y. Zhang, and D. Yao, “Missing data imputation for traffic flow based on improved local least squares,” *Tsinghua Science and Technology*, vol. 17, no. 3, pp. 304–309, 2012.
 - [23] H. T. Yang, J. J. Yang, L. D. Han et al., “A Kriging based spatiotemporal approach for traffic volume data imputation,” *PLoS One*, vol. 13, no. 4, article e0195957, 2018.
 - [24] V. Audigier, F. Husson, and J. Josse, “Multiple imputation for continuous variables using a Bayesian principal component analysis,” *Journal of Statistical Computation and Simulation*, vol. 86, no. 11, pp. 2140–2156, 2016.
 - [25] Y. Wang, Y. Zhang, X. L. Piao, H. Liu, and K. Zhang, “Traffic data reconstruction via adaptive spatial-temporal correlations,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 4, pp. 1531–1543, 2019.
 - [26] G. Pastor, “A low-rank tensor model for imputation of missing vehicular traffic volume,” *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8934–8938, 2018.
 - [27] Y. X. Han and Z. C. He, “Simultaneous incomplete traffic data imputation and similarity pattern discovery with Bayesian nonparametric tensor decomposition,” *Journal of Advanced Transportation*, vol. 2020, Article ID 8810753, 2020.
 - [28] X. Kong, K. Wang, S. Wang et al., “Real-time mask identification for COVID-19: an edge computing-based deep learning framework,” *IEEE Internet of Things Journal*, vol. 8, no. 21, pp. 15929–15938, 2021.
 - [29] A. Montieri, G. Bovenzi, G. Aceto, D. Ciunzio, V. Persico, and A. Pescapè, “Packet-level prediction of mobile-app traffic using multitask deep learning,” *Computer Networks*, vol. 200, article 108529, 2021.
 - [30] M. Liang, R. W. Liu, S. C. Li, Z. Xiao, X. Liu, and F. Lu, “An unsupervised learning method with convolutional auto-encoder for vessel trajectory similarity computation,” *Ocean Engineering*, vol. 225, article 108803, 2021.
 - [31] Y. Duan, Y. Lv, Y. L. Liu, and F. Y. Wang, “An efficient realization of deep learning for traffic data imputation,” *Transportation Research Part C: Emerging Technologies*, vol. 72, pp. 168–181, 2016.
 - [32] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: feature learning by inpainting,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2536–2544, Las Vegas USA, 2016.
 - [33] J. Yoon, J. Jordon, and M. Van Der Schaar, “GAIN: missing data imputation using generative adversarial nets,” in *Proceedings of 35th International Conference on Machine Learning*, Stockholm, Sweden, 2018.
 - [34] Y. Y. Chen, Y. S. Lv, and F. Y. Wang, “Traffic flow imputation using parallel data and generative adversarial networks,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 4, pp. 1624–1630, 2020.
 - [35] A. Kazemi and H. Meidani, *IGANI: iterative generative adversarial networks for imputation applied to prediction of traffic data*, 2020, <https://arxiv.org/abs/2008.04847v1>.
 - [36] G. J. Shen, Z. Z. Zhao, and X. J. Kong, “GCN2CDD: a commercial district discovery framework via embedding space

- clustering on graph convolution networks,” *IEEE Transactions on Industrial Informatics*, vol. 18, no. 1, pp. 356–364, 2022.
- [37] X. Han, G. J. Shen, X. Yang, and X. J. Kong, “Congestion recognition for hybrid urban road systems via digraph convolutional network,” *Transportation Research Part C:Emerging Technologies*, vol. 121, article 102877, 2020.
- [38] Y. X. Chen and Z. C. He, “Vehicle identity recovery for automatic number plate recognition data via heterogeneous network embedding,” *Sustainability*, vol. 12, no. 8, p. 3074, 2020.
- [39] Z. Y. Cui, L. F. Lin, Z. Y. Pu, and Y. H. Wang, “Graph Markov network for traffic forecasting with missing data,” *Transportation Research Part C:Emerging Technologies*, vol. 117, article 102671, 2020.
- [40] J. X. You, X. B. Ma, D. Y. Ding, M. Kochenderfer, and J. Leskovec, *Handling missing data with graph representation learning*, 2020, <https://arxiv.org/abs/2010.16418v1>.
- [41] A. R. Benson, D. F. Gleich, and J. Leskovec, “Higher-order organization of complex networks,” *Science*, vol. 353, no. 6295, pp. 163–166, 2016.
- [42] J. Piao, G. Zhang, F. Xu, Z. Chen, and Y. Li, “Predicting customer value with social relationships via motif-based graph attention networks,” in *Proceedings of the Web Conference*, Ljubljana, Slovenia, 2021.
- [43] A. M. Haasnoot, M. W. Schilham, S. Kamphuis et al., “Identification of an amino acid motif in HLA-DR1 that distinguishes uveitis in patients with juvenile idiopathic arthritis,” *Arthritis & Rheumatology*, vol. 70, no. 7, pp. 1155–1165, 2018.
- [44] Y. Zhu, A. Mueen, and E. Keogh, “Matrix profile IX: admissible time series motif discovery with missing data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 6, pp. 2616–2626, 2021.
- [45] T. H. Do, D. M. Nguyen, and G. Bekoulis, *Graph convolutional neural networks with node transition probability-based message passing and drop node regularization*, 2020, <https://arxiv.org/abs/2008.12578>.
- [46] N. Zhang, X. F. Guan, J. Cao, X. L. Wang, and H. Y. Wu, “Wavelet-HST: a wavelet-based higher-order spatio-temporal framework for urban traffic speed prediction,” *IEEE Access*, vol. 7, pp. 118446–118458, 2019.
- [47] W. Ma, L. Cai, T. He, L. Chen, Z. Cao, and R. Li, “Local expansion and optimization for higher-order graph clustering,” *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8702–8713, 2019.
- [48] G. Aceto, D. Ciunzo, A. Montieri, and A. Pescape, “Mobile encrypted traffic classification using deep learning: experimental evaluation, lessons learned, and challenges,” *IEEE Transactions on Network and Service Management*, vol. 16, no. 2, pp. 445–458, 2019.
- [49] K. Zhang, X. H. Zhao, X. Li, X. Y. You, and Y. H. Zhu, “Network traffic prediction via deep graph-sequence spatiotemporal modeling based on mobile virtual reality technology,” *Wireless Communications & Mobile Computing*, vol. 2021, article 2353875, 2021.
- [50] F. Shahid, A. Zameer, and M. Muneeb, “Predictions for COVID-19 with deep learning models of LSTM, GRU and bi-LSTM,” *Chaos Solitons & Fractals*, vol. 140, article 110212, 2020.
- [51] R. W. Liu, J. Nie, S. Garg, Z. Xiong, Y. Zhang, and M. S. Hosain, “Data-driven trajectory quality improvement for promoting intelligent vessel traffic services in 6G-enabled maritime IoT systems,” *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5374–5385, 2021.