IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 18, 2020, accepted March 2, 2020, date of publication March 5, 2020, date of current version March 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2978530

Traffic Data Imputation and Prediction: An Efficient Realization of Deep Learning

JUNHUI ZHAO[1,2], (Senior Member, IEEE), YIWEN NIE[2], SHANJIN NI[3], AND XIAOKE SUN[2]

[1]School of Information Engineering, East China Jiaotong University, Nanchang 330013, China
[2]School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China
[3]National Computer Network Emergency Response Technical Team/Coordination Center of China (CNCERT/CC), Beijing 100029, China

Corresponding author: Junhui Zhao (junhuizhao@hotmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61661021 and 61971191, in part by the National Science and Technology Major Project of the Ministry of Science and Technology of China under Grant 2016ZX03001014-006, in part by the Open Research Fund of the National Mobile Communications Research Laboratory, Southeast University, under Grant 2017D14, and in part by the Beijing Natural Science Foundation under Grant L182018.

ABSTRACT In this paper, we study the prediction of traffic flow in the presence of missing information from data set. Specifically, we adopt three different patterns to model the missing data structure, and apply two types of approaches for the imputation. In consequence, a forecasting model via deep learning based methods is proposed to predict the traffic flow from the recovered data set. The experiments demonstrate the effectiveness of using deep learning based imputation in improving the accuracy of traffic flow prediction. Based on the experimental results, we conduct a thorough discussion on the appropriate methods to predict traffic flow under various missing data conditions, and thus shedding the light for a practical design.

INDEX TERMS Data missing imputation, deep learning, traffic flow prediction.

I. INTRODUCTION

Intelligent Transportation systems (ITSs), which aims at providing innovative services and making safer, more convenient use of traffic network, typically depend on traffic flow information, i.e., the number of vehicles crossing a specific region per unit time interval, as inputs to make the underlying decision logic [1], [2]. As such, accurate and timely traffic flow information is critical for engineers and researchers to assess the performance of traffic system, relieve the traffic congestion, and improve the traffic efficiency. Moreover, to help operators response in a more adequate way, the prediction of traffic flow has been introduced [3], whereas the short term future traffic flow can be estimated from historical observations. Especially as the vital vehicular technology for realization of automatic vehicle driving, congestion control, and the emerging cellular vehicle-to-everything (C-V2X) based smart city [4], [5], such topic has attracted significant attention followed by a variety of studies, including time-series approaches [6], probabilistic graph approaches [7], and nonparametric approaches such as artificial neural networks (ANNs) [8].

In the domain of ANNs based methods, deep learning is considered as one of the most effective and efficient traffic flow prediction approaches. Due to the nonlinear nature of traffic flow information, the conventional parameter approaches cannot exactly capture the traffic flow features with analysis formulas. However, the nonlinearity and randomness issues can be addressed well by deep learning theory. Compared with the traditional shallow learning architectures, deep neural network is able to model deep complex non-linear relationship by using distributed and hierarchical feature representation [3]. Currently, deep learning has made certain progress in the domains including speech recognition, computer vision, and natural language processing. Under the guide of deep learning theory, many neural network variants have been proposed to solve the traffic forecasting problem [9].

However, the prediction of traffic flow relies heavily on the complete data set of historical observations collected from a variety of senor sources. With the densely deployed traffic sensors and the new emerging sensor techniques in the fifth

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

generation mobile communication (5G) era, the amount of traffic data is exploding [11]. The availability of huge amount of traffic data has the potential to lead to a great revolution in the development of ITSs. It is foreseeable that the future ITSs will become a more powerful multifunctional data-driven intelligent transportation system.

Missing data usually occurs in the real world settings and thus limiting the performance of predicting methods [12]. To this end, it is key to handle the missing data appropriately before applying any machine learning algorithms. Fortunately, in most cases, the attributes of traffic flow data are intertwined with each other [13], [14]. According to the identification of relationship among attributes, the values of traffic data missing points can be subsequently determined. Up to day, there have been many studies in the literature regarding data missing imputation [15], [16]. The authors in [17] illustrated that the accuracy of prediction could be increased by the imputation methods in the presence of missing-data perturbation. For instance, by extracting the nonlinear cross-correlation features involved in the missing data, an NN-based imputation method was proposed to estimate the missed observation [15]. The k-Nearest Neighbors (k-NN) algorithm was used in [12] as an imputation method to produce some plausible values that could be used to replace the missing values in a data set. By using the correlations contained in the traffic data structure, deep learning based approaches for traffic data imputation were proposed in [16], [18]. The authors in [16] illustrated that deep learning could be applied in the field of data imputation. A Generative Adversarial Networks (GAN) based imputation method was proposed in [18].

Nevertheless, most previous works only considered the problem of data missing imputation existed in the traffic data, and did not consider the traffic flow prediction problem in the presence of missing information from traffic data sets [1], [19]–[21].

In this paper, we target at predicting the traffic flow using data sets with missing-element perturbation. Specifically, to account for that the missing data can occur with various patterns, we analysis four different practical structures and adopt three models of them, i.e., data missing at random (DMR), block data missing (BDM), and multi-block data missing (MBDM). Based on these data missing models, we present two different methods for the imputation, including mean imputation and deep learning based imputation. Three deep learning methods including Stacked Autoencoders (SAEs), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) are then used to extract generic traffic flow features for prediction with the filled traffic data. The effectiveness of the proposed approaches are demonstrated via the inference experiments. Subsequential discussions are also provided. The major contribution of our paper can be summarized as follows:

- We consider the traffic flow prediction problem in the presence of missing data imputation with a thorough and practical analysis of patterns of missing data. Different from other traffic flow prediction methods, the proposed scheme can not only effectively solve the traffic data missing problem occurred in the real world settings, but also obtain the accurate prediction results of traffic flow.
- Deep learning based imputation and prediction (DLIP) model is proposed to improve the accuracy of traffic flow prediction under missing data simultaneously. The obtained experiment results have verified the forecasting performance of the proposed scheme. For the normal scenarios with less than 40 percent data missing, deep learning based imputation methods like SAEs, LSTM and GRU can achieve less error of the prediction than the traditional ones. Furthermore, the flexibility of the DLIP model comes from the fact that any state-of-the-art deep learning based method can be utilized into the model to enhance the system performance on prediction and imputation.

The rest of the paper is organized as follows. In Section II, the methodology of traffic flow imputation and prediction is introduced. In Section III, we describe our data missing imputation models. The performance of the proposed scheme is evaluated in Section IV before concluding the paper in Section V.

II. METHODOLOGY

In this section, we introduce the traffic flow prediction problem, deep learning based methods, the general training procedure, and the proposed model for imputation and prediction.

A. TRAFFIC FLOW PREDICTION PROBLEM

The prediction of traffic flow requires not just a sufficiently large collection of recorded data, but more importantly, a reliable mechanism that extracts the intrinsic structure from the data set and predicts the future flow. To facilitate this task, we first structure the collected data set in the form of sample vectors $\mathbf{x} = \{x_1, \ldots, x_t, \ldots, x_T\}$, where x_t denotes the observed traffic flow quantity during the t^{th} time interval. Next, we leverage an unsupervised learning technique, that adopts the autoencoder blocks to create deep networks and is referred to as SAEs [1], for the predictor design. Particularly, let $f_1, \ldots, f_n$ denote the univariate activation link functions, e.g., the sigmoid function, at each layer. The predictor $\hat{y}$ at the output is then defined as a series of functional composition, i.e.,

$$\hat{y}(x) := (f_n \circ \ldots \circ f_1)(x), \tag{1}$$

whereas f_l is a semi-activation rule [19], given by

$$\begin{aligned} f_l(x) &= f\left(\sum_{j=1}^{N_l} w_{lj} x_j + z_j\right) \\ &= f\left(w_l^T x_l + z_l\right), \quad l = 1, \ldots, n, \end{aligned} \tag{2}$$

where N_l is the number of units at layer l, $w = [w_1, \ldots, w_n]$ is a weight matrix, and $z = [z_1, \ldots, z_n]$ is a bias vector.

To this end, we can formally define the traffic flow prediction problem as follows:

Problem Formulation: Given the input traffic flow data $x_t = \{x_1, x_2, \ldots, x_t\}$, the problem is to use (1) to predict the traffic flow at the time interval $(t + \Delta)$ for some prediction horizon Δ by learning the weights $w_l \in \mathbb{R}^{N_l \times N_{l-1}}$ and bias $z_l \in \mathbb{R}^n$.

B. DEEP LEARNING BASED METHODS

1) SAEs

A basic autoencoder is a neural network, in which an original signal x at the input is reproduced based on the reconstructed error between the input x and the network's output y. The autoencoder model tries to learn the compressed and distributed features in hidden nodes and reproduce the input signals at the output of the model, which consists of one input layer, one hidden layer, and one output layer.

An autoencoder first encodes an input x to a hidden representation $h(x)$, and the decoding is then processed, which the representation $h(x)$ is backward expressed from the hidden layer to input layer. The encoding and decoding process can be presented as

$$\begin{aligned} h &= f_e\left(wx + b\right), \\ y &= f_d\left(\tilde{w}x + c\right), \end{aligned} \tag{3}$$

where $f_e(x)$ denotes the encoding activation function, and $f_d(x)$ denotes the decoding activation function. In this paper, the logistic sigmoid function

$$f_{sigm}(x) = \frac{1}{1 + e^{-x}}, \tag{4}$$

is considered for $f_e(x)$ and $f_d(x)$, and the weight matrix $\tilde{w} = w^T$ is assumed.

The reconstruction y can be considered as the predictive value of x, and in order to get the exactly $\theta = \{w, b, c\}$, the reconstruction error $\mathcal{L}(x, y)$ needs to be minimized, i.e.,

$$\theta = \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}(x, y) = \underset{\theta}{\operatorname{argmin}}\, \|x - y\|^2. \tag{5}$$

The SAEs [22] can be defined via an extension of the above concept. In particular, an SAEs model consists of several hidden layers, whereas the output of a generic hidden layer is used as the input of the next hidden layer. For instance, in an SAE with l layers, an autoencoder is applied to train the first layer with the training data as inputs, a standard predictor is added on the last second layer to predict the short-term traffic flow, the last output layer is used to output the predicted data. Moreover, the hidden layers are used to learn the abstract features, and the output of the i^{th} hidden layer is used as the input of the $(i+1)^{th}$ hidden layer. The classic construction of SAEs model is illustrated in Fig. 1.

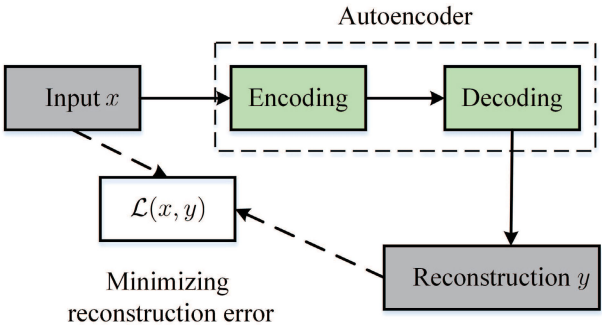


FIGURE 1. The construction of SAEs model.

2) LSTM

Except for the typical deep neural network like SAEs, recent examples applied in ITSs include convolutional neural network (CNN) [23], recurrent neural network (RNN) [24], and GAN [18], [25]. It is worth mentioning that the RNN is widely adopted as a suitable method to capture the temporal and spatial evolution of traffic flow. However, previous studies proved that the traditional RNN failed to capture the long-term evolution because of some existed challenges including vanishing gradient and exploding gradient [26].

To address these problems, the novel structures of RNNs like LSTM and GRU were proposed with mechanism of information gates, which was designed to give the memory cells ability to determine when to forget certain information. The vital change of network structure design creates the optimal time lags for RNNs. LSTM network was proposed [27] and applied well in short-term traffic forecasting [28]. Compared with the conventional RNNs, LSTM has the ability to capture the features of time series within longer time span. Therefore, the traffic forecasting can achieve a better performance via the LSTM network.

The simplified architecture of LSTM methods shown in Fig. 2(a) provides a better understanding. It is illustrated here for the propagation of hidden states h among the neural networks. There are three gates constituting a common LSTM cell as described below. The cell remembers values over arbitrary time intervals and the gates regulate the flow of information into and out of the cell. We denote W_x and W_h as the weight parameters of the current input information x_t and the hidden state vector h_t from the neural network, respectively.

- *Forget-gate:* It determines whether x_t should be retained or not. Information from the previous hidden state h_{t-1} and from x_t is passed through the sigmoid function. The output value of the function ranges from 0 to 1 as the forgetting vector f_t to make decisions for filtering nonsignificant information, which can be written as

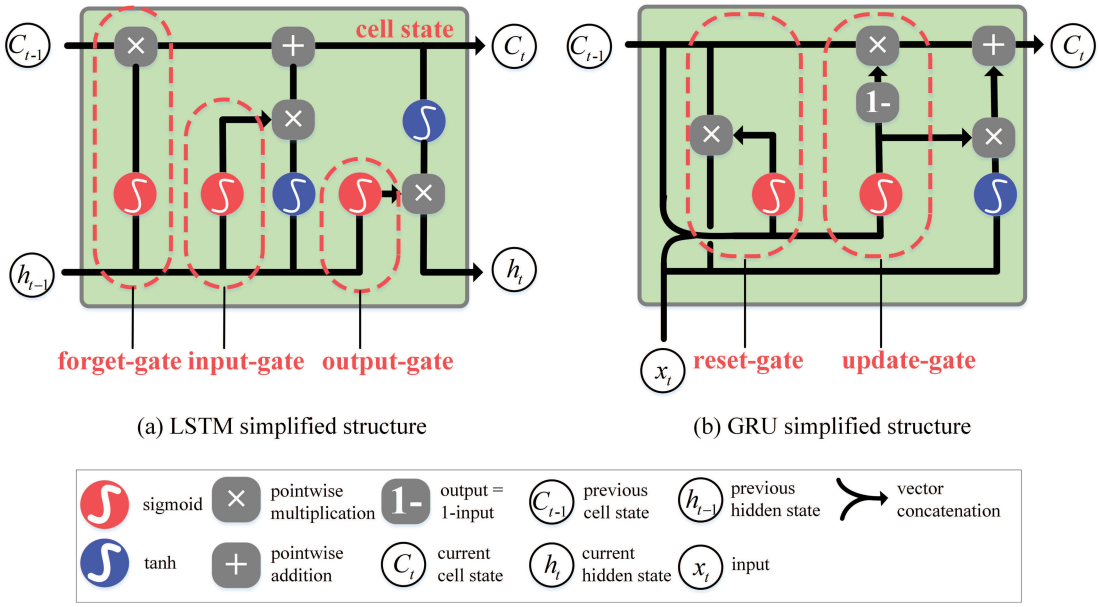
$$f_t = f_{sigm}\left(W_{xf}x_t + W_{hf}h_{t-1}\right). \tag{6}$$

- *Input-gate:* The processed value of the input-gate can be expressed as

$$i_t = f_{sigm}\left(W_{xi}x_t + W_{hi}h_{t-1}\right). \tag{7}$$

  Then the historical and the current information are input into a hyperbolic tangent (tanh) function, which can be written as

$$\widetilde{c}_t = f_{tanh}\left(W_{xc}x_t + W_{hc}h_{t-1}\right). \tag{8}$$



(a) LSTM simplified structure (b) GRU simplified structure

FIGURE 2. The simplified architecture of LSTM and GRU methods.

The tanh function is a typical activation function, given by

$$f_{tanh}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \tag{9}$$

It compresses the input data into the range $[-1,1]$ for outputting the preprocessed input value of cell state. The current state value c_t is updated as

$$c_t = f_t \odot c_{t-1} + i_t \odot \widetilde{c}_t, \tag{10}$$

and propagated into the next LSTM cell, where the operator $\odot$ represents pointwise multiplication.

- *Output-gate:* The output value is the result of the output-gate, given by

$$o_t = f_{sigm}\left(W_{xo}x_t + W_{ho}h_{t-1}\right). \tag{11}$$

Finally, the current hidden state is obtained as

$$h_t = o_t \odot f_{tanh}\left(c_t\right). \tag{12}$$

3) GRU

As a significant variant algorithm of LSTM, the GRU [29] shows the comprehensive abilities in time-series forecasting problems via the combination of the forget-gate and the input-gate into the reset-gate.

As shown in Fig. 2(b), the update-gate is in charge of inputting and discarding information, which covers the work of the input-gate and the forget-gate in LSTM. The reset-gate focuses on how much previous information to be discarded. In GRU, the fewer parameters are computed and processed, and the hidden state is propagated directly among the network cells instead of being controlled by the output-gate. Thus, GRU is similar to LSTM, but simpler to compute and implement for speeding up the procedure of training.

C. TRAINING PROCEDURE

Training the traffic flow data set and extracting the features for the prediction are the key process of the experiment. General training procedures of deep learning based methods contain the inputting data, the updating parameters of the natural network by minimizing the loss function, and the outputting the results. Take SAEs as an example, the training procedure can be concluded as follows:

1) Given training sets x and the number of hidden layers l;
2) Train the first layer as an autoencoder by minimizing $\mathcal{L}(x, y)$ with x as the input;
3) Use the output of the first layer as the input, and train the second layer as an autoencoder;
4) Iterate as in 3) to the desired number of hidden layers l;
5) Predict the traffic flow with the output of the hidden layer as the input of the prediction layer;
6) Fine-tune the parameters of the whole network.

D. THE PROPOSED MODEL

The proposed DLIP model for traffic flow data imputation and prediction is shown in Fig. 3. As the input matrices of imputation model, the traffic flow data on three different missing models are first fed into the deep natural network and the traditional methods, respectively. After the mentioned training procedure using three different deep learning based methods and the computation via conventional methods, the completed data sets are input into the prediction model to obtain the forecasting results by using the traditional methods and the deep learning based methods, respectively. However, it should be noted that the prediction model is obtained by training of a extra complete traffic flow data. The predicted data are evaluated by the prediction and imputation errors.

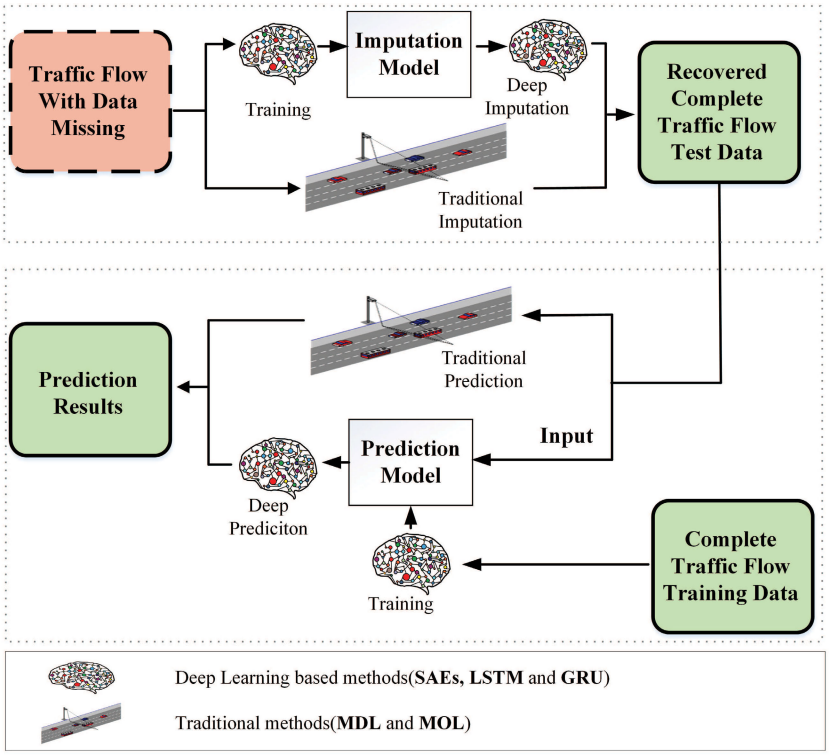


FIGURE 3. The proposed DLIP model with deep learning based methods for traffic flow imputation and prediction.

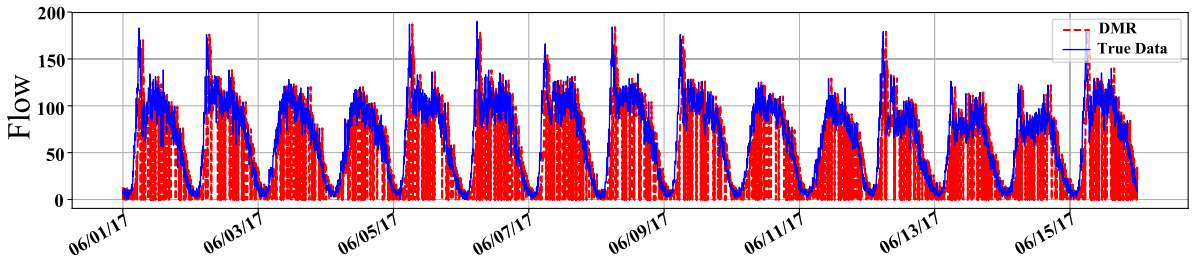


FIGURE 4. Traffic flow with data missing on DMR model.

III. DATA MISSING IMPUTATION

In practice, data collected from the ground loop detectors and video surveillance cameras are often corrupted or with certain content missing due to device variation. As many machine learning models rely on the complete data sets, it is often required that the missing data shall be imputed prior to running statistical analysis [16]. In this section, we introduce three different structures, i.e., DMR, BDM, and MBDM, to model the missing data [30]. Note that such framework can readily incorporate various data missing patterns. Besides, we also propose two different types of methods for the data imputation task.

A. DATA MISSING MODELS

In consideration to a practical scenario, we introduce three different patterns to model the missing data set, according to [31].

- *DMR Model*: The missing data occurs in a point-wise pattern according to a universal constant probability, shown in Fig. 4. Traffic flow data miss in a random way and it is completely independent upon the missing values.
- *BDM Model*: In this case, a consecutive trunk of data points, refer to as block, disappear simultaneously, shown in Fig. 5. The missing data distributes in the data set in a single block way.
- *MBDM Model*: Different from the BDM, in this case, the missing data occurs in several none overlapping blocks, shown in Fig. 6. The missing data in the MBDM model are divided into many small blocks, and these blocks distribute in the whole data set randomly.

Actually, another pattern called missing at non-random (MANR) due to the long-term data detector failure may exist in the traffic flow data set. On the special pattern, there is some correlation between the missing data and the characteristics of the missing data itself. The location of the missing data depends on both the missing data itself and other missing data. Consistent with previous studies [32], we assume

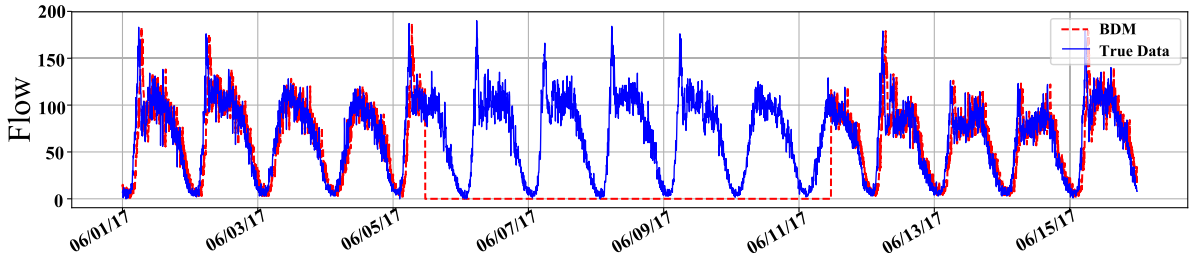


FIGURE 5. Traffic flow with data missing on BDM model.

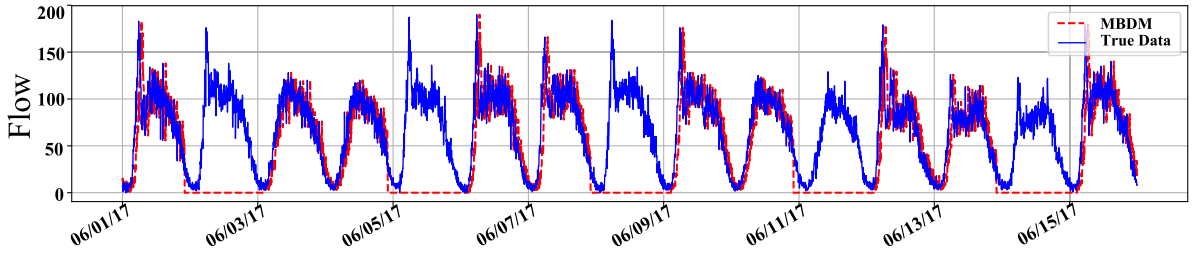


FIGURE 6. Traffic flow with data missing on MBDM model.

that non-random missing data have been discovered and deleted.

Among all the cases, we denote the proportion of missing data in the complete data set as missing rate.

B. DATA MISSING IMPUTATION METHODS

There are many imputation methods to fill the values of data missing points, which range from the simple mean imputation method to the complex deep learning methods based on the intrinsic properties of data set. In this paper, we introduce some imputation methods, given as

1) *Traditional imputation*: Conventional methods including mean imputation, nearest time distance [20], and ARIMA models [33] made some progress in short-term traffic flow forecasting problem. Considering the sensibility for time delay in the ITSs, the lower algorithm complexity, and the better universality, we choose the mean method to fit the scenario of traffic flow imputation and prediction.

   In general, such method recovers the missing data by using the mean of other available data. Specifically, with the traffic data obtained from M lanes, we propose two mean imputation methods. The first mean imputation method, coined as the mean of desired lane (MDL), only depends on the traffic data of the desired lane, whereas the filled value $\hat{x}_d^{a_i=0}$ can be expressed as

$$\hat{x}_d^{a_i=0} = \frac{1}{N - N_z} \sum_{i=1}^{N} a_i x_i, \quad a_i \in \{0, 1\}, \tag{13}$$

   with $x = [x_1, \ldots, x_i, \ldots, x_N]$ being the whole data set containing the data missing points, a_i indicates whether the value of the i^{th} data is available ($a_i = 1$) or not ($a_i = 0$), whereas N and N_z denote the number of total data variables and the number of data missing points, respectively. Our second imputation method, termed the mean of other lanes (MOL), utilizes the correlation of traffic condition among different lanes to impute the missing data. Without loss of generality, let the k^{th} lane be the target lane, the imputation value $\hat{x}_o^j$ can be expressed as

$$\hat{x}_o^j = \frac{1}{M - 1} \sum_{m=1, m\neq k}^{M} x_{mj}, \tag{14}$$

   where j is the index of data missing point in k^{th} lane. However, due to the non-linear and stochastic nature of traffic flow, these conventional methods may not abstract the unique nature well, which results in the bigger prediction errors than the deep learning based methods.

2) *Deep learning based imputation*: With this approach, we apply more advanced deep learning based techniques, i.e., the SAEs, LSTM, and GRU methods, to deal with the data missing problem. The prediction of traffic flow heavily relies on the complete data set of historical observations, and the deep learning based methods have the capability to extract intrinsic features from the traffic flow with data missing to get a more precise prediction. Firstly, the complete data $x_{mi}, m \in [1, M]$ is regarded as the training data, once the training model is established, the data x_{kj} ($a_j = 0, m \neq k$) of other lanes is then used to predict the value of data missing point in k^{th} lane based on the training model. The deep learning based data missing imputation algorithm is summarized in Algorithm 1. In Algorithm 1, the traffic flow data set with data lost on three different models is fed into the network algorithm. The first loop is for obtaining the index of missing data. Next, the normalization is conducted by preprocessing the processed data set without missing data. Then, the normalized

Algorithm 1 Learning Methods Based Missing Data Imputation Algorithm

Input: Traffic flow data set x with data lost
Output: Data set $\tilde{x}$ with missing data filled
1: Step 1) Obtain the index of missing data
2: **for** i in $[0, \text{length}(x) - 1]$ **do**
3: **if** $x[i]$ is 0 **then**
4: index $= i$;
5: Delete the missing data, and get the processed data set $\hat{x}$;
6: **end if**
7: **end for**
8: Step 2) Preprocess the data set $\hat{x}$
9: **for** x in $\hat{x}$ **do**
10: $x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$;
11: **end for**
12: Obtain the training data set $\bar{x} = \text{processed}(\hat{x})$;
13: Step 3) Train model
14: Train the network with learning method, and get the training model;
15: Step 4) Fill the missing data
16: Predict the missing data by using the obtained training model, and obtain the filled data set $\tilde{x}$.
17: **return** $\tilde{x}$

data $\bar{x}$ is trained by the deep learning based methods, i.e., SAEs, LSTM, and GRU. Finally, the missing data is filled by the training of the deep neutral network and the filled data set $\tilde{x}$ is obtained.

IV. EXPERIMENTS

In this section, we conduct a series of experiments on the traffic flow data set taken from Caltrans Performance Measurement System (PeMS) database [34], and evaluate the effectiveness of our proposed methods through a variety of performance metrics.

A. DATASET DESCRIPTION

The experimental time-series data set used for the imputation and prediction of traffic flow is collected from the PeMS database. PeMS started in 1999, and it provides a consolidated database of traffic data collected every 30s from over 35,000 detectors, which are placed on state highways throughout California. The collected traffic data are aggregated into 5-minute increments for each detector station. The traffic data used for the experiments were collected in three months (i.e., April, May, and June) of the year 2017. The traffic data of April and May are selected as the training data, and the data of June are selected as the test data.

B. INDEX OF PERFORMANCE

In order to evaluate the effectiveness of the imputation methods, as well as the accuracy of traffic flow prediction, we introduce three different metrics, i.e., the mean absolute error (MAE), the mean relative error (MRE), and the normalized mean square error (NMSE). Formally, these metrics are respectively defined as follows

$$\begin{aligned} \text{MAE} &= \frac{1}{n}\sum_{i=1}^{n} |x_i - \hat{x}_i|, \\ \text{MRE} &= \frac{1}{n}\sum_{i=1}^{n} \frac{|x_i - \hat{x}_i|}{x_i}, \\ \text{NMSE} &= \frac{\sum_{i=1}^{n} |x_i - \hat{x}_i|^2}{\sum_{i=1}^{n} x_i^2}, \end{aligned} \tag{15}$$

where x_i is the original traffic data, and $\hat{x}_i$ is the filled or predicted traffic data.

TABLE 1. Structure of deep learning based methods.

Training Model	Imputation		Prediction	
	Hidden Layers	Hidden Units	Hidden Layers	Hidden Units
SAEs	3	[100,100,100]	3	[400,400,400]
LSTM	2	[64,64]	2	[64,64]
GRU	2	[64,64]	2	[64,64]

C. PARAMETER SETTING AND PERFORMANCE EVALUATION

In this paper, we use the traffic flow data at time interval t from other lanes to interpret the value of data missing point at time stamp t from the desired lane. As for the traffic flow prediction, we use the traffic data among previous r time intervals to predict the traffic flow at time interval t based on the spatial correlations of traffic flow, i.e., using $x_{t-1}, x_{t-2}, \ldots, x_{t-r}$ to predict x_t.

In our inference experiment, the test data are chosen as the data missing set due to the fact that the data missing problem in the training data set with such rich amount of traffic data will not greatly affect the accuracy of extracting the features of traffic data, while the test data with small amount of data will be greatly influenced by the data missing problem. We use the complete test data set, and the values of these data are replaced by zeros based on the data missing models and data missing rates. The number of lanes is $M = 5$, and we use the DLIP model to predict the 60-min traffic flow, where sufficient data set is needed for training.

In the deep learning based data missing imputation methods, we set the number of time intervals $r = 4$, and in the traffic flow prediction, we set $r = 12$. The structure of deep learning based data missing imputation and traffic flow prediction methods are given in Table 1. The mentioned hyper parameters are designed based on the performance in the simulations.

In Table 2, we compare the prediction accuracy of different deep learning based traffic flow prediction methods under the DMR model. We observe that the prediction methods with deep learning based imputation achieve a better performance than those using mean imputation. This is mainly because the predicted values of deep learning based imputation methods are much closer to the true values than that of mean imputation methods. Moreover, among the prediction methods with deep learning based imputation, we can clearly see that the prediction method with GRU-based imputation attains a much better performance than other methods.

TABLE 2. Comparison of prediction accuracy for different methods based on DMR model.

Imputation Methods	Training Model	15%			20%			25%			30%		
		MAE	MRE	NMSE	MAE	MRE	NMSE	MAE	MRE	NMSE	MAE	MRE	NMSE
MDL	SAEs	11.25	54.97%	3.65%	12.42	65.33%	4.35%	13.38	70.47%	5.18%	14.51	85.98%	5.93%
	LSTM	11.25	55.72%	3.78%	12.42	66.10%	4.48%	13.44	70.46%	5.37%	14.52	84.87%	6.05%
	GRU	11.92	67.75%	4.48%	13.02	79.46%	5.20%	13.97	84.55%	6.10%	15.03	99.81%	6.82%
MOL	SAEs	9.04	19.72%	2.49%	9.57	20.03%	2.84%	10.26	20.59%	3.29%	10.93	21.16%	3.78%
	LSTM	9.37	18.93%	2.70%	10.09	19.62%	3.15%	10.88	20.44%	3.67%	11.75	21.25%	4.31%
	GRU	8.98	19.13%	2.50%	9.57	19.64%	2.88%	10.18	20.16%	3.26%	10.88	20.89%	3.75%
SAEs	SAEs	8.01	18.64%	1.87%	7.97	18.60%	1.88%	8.23	18.41%	1.99%	8.30	18.34%	2.04%
	LSTM	8.04	17.54%	1.89%	7.97	17.20%	1.87%	8.56	18.33%	2.10%	8.60	18.08%	2.14%
	GRU	7.91	18.06%	1.83%	7.79	17.77%	1.81%	8.14	18.28%	1.93%	8.19	18.19%	1.95%
LSTM	SAEs	8.02	18.55%	1.86%	8.03	18.98%	1.90%	8.03	18.37%	1.88%	8.18	18.44%	1.99%
	LSTM	8.18	17.89%	1.91%	8.00	17.31%	1.90%	8.24	18.01%	1.96%	8.48	17.97%	2.10%
	GRU	7.99	18.21%	1.84%	7.86	18.03%	1.84%	7.98	18.23%	1.84%	8.07	18.10%	1.94%
GRU	SAEs	7.94	19.39%	1.82%	7.99	18.53%	1.88%	7.97	18.54%	1.88%	8.06	18.55%	1.95%
	LSTM	7.83	17.35%	1.78%	8.07	17.41%	1.90%	8.02	17.36%	1.89%	8.25	17.60%	2.02%
	GRU	7.78	18.25%	1.77%	7.86	17.95%	1.82%	7.84	17.91%	1.82%	7.92	17.97%	1.88%

TABLE 3. Comparison of prediction accuracy for different methods based on BDM model.

Imputation Methods	Training Model	15%			20%			25%			30%		
		MAE	MRE	NMSE	MAE	MRE	NMSE	MAE	MRE	NMSE	MAE	MRE	NMSE
MDL	SAEs	12.07	52.60%	5.68%	13.50	66.44%	6.93%	14.51	72.26%	7.82%	15.62	84.88%	8.95%
	LSTM	12.27	53.46%	5.80%	13.65	65.98%	6.94%	14.69	71.30%	7.81%	15.78	82.75%	8.79%
	GRU	12.19	47.58%	5.86%	13.58	60.25%	6.96%	14.64	65.64%	7.84%	15.72	77.27%	8.77%
MOL	SAEs	10.84	23.16%	4.47%	11.82	24.34%	5.26%	12.87	25.52%	6.14%	13.34	26.11%	6.43%
	LSTM	11.15	27.06%	4.80%	12.14	28.17%	5.61%	13.24	29.33%	6.55%	13.75	30.02%	6.87%
	GRU	11.33	21.23%	5.17%	12.49	22.85%	6.14%	13.74	24.45%	7.25%	14.37	25.31%	7.66%
SAEs	SAEs	8.15	20.31%	2.02%	8.23	20.57%	2.05%	8.76	20.67%	2.31%	8.51	20.24%	2.24%
	LSTM	8.35	24.23%	2.11%	8.50	24.61%	2.17%	9.14	24.32%	2.53%	8.98	24.23%	2.46%
	GRU	8.16	17.63%	2.05%	8.32	17.95%	2.11%	9.21	19.43%	2.56%	9.04	19.03%	2.48%
LSTM	SAEs	8.27	21.04%	2.09%	8.39	21.45%	2.14%	8.89	20.89%	2.37%	8.79	20.60%	2.33%
	LSTM	8.52	25.08%	2.22%	8.73	25.68%	2.30%	9.27	24.37%	2.59%	9.27	24.37%	2.59%
	GRU	8.29	17.96%	2.15%	8.52	18.39%	2.24%	9.37	19.78%	2.64%	9.42	19.75%	2.64%
GRU	SAEs	8.30	20.90%	2.12%	8.48	20.51%	2.14%	8.78	22.00%	2.38%	8.98	21.04%	2.42%
	LSTM	8.57	24.96%	2.26%	8.76	24.21%	2.29%	9.29	26.48%	2.65%	9.49	24.98%	2.70%
	GRU	8.37	17.99%	2.20%	8.75	18.79%	2.29%	9.16	19.23%	2.63%	9.66	20.14%	2.77%

TABLE 4. Comparison of prediction accuracy for different methods based on MBDM model.

Imputation Methods	Training Model	15%			20%			25%			30%		
		MAE	MRE	NMSE	MAE	MRE	NMSE	MAE	MRE	NMSE	MAE	MRE	NMSE
MDL	SAEs	13.86	80.71%	7.96%	14.89	79.14%	8.63%	16.07	77.75%	9.44%	16.74	77.16%	9.86%
	LSTM	13.91	82.97%	7.83%	14.99	82.07%	8.60%	16.15	81.31%	9.43%	16.82	81.07%	9.84%
	GRU	13.70	77.25%	7.61%	14.88	76.39%	8.51%	16.16	75.69%	9.49%	16.92	75.53%	10.00%
MOL	SAEs	9.76	20.17%	3.75%	11.80	22.14%	5.77%	13.55	23.86%	7.32%	14.80	25.25%	8.27%
	LSTM	9.76	24.38%	3.78%	11.66	26.22%	5.60%	13.30	27.84%	7.02%	14.46	29.15%	7.86%
	GRU	9.93	20.30%	3.95%	11.97	22.28%	6.01%	13.73	24.03%	7.62%	15.04	25.53%	8.63%
SAEs	SAEs	9.48	19.51%	3.62%	10.96	21.55%	4.40%	10.26	20.70%	3.34%	15.86	27.26%	9.49%
	LSTM	9.44	23.44%	3.59%	10.92	25.25%	4.45%	10.51	24.93%	3.62%	15.30	29.58%	8.77%
	GRU	9.62	19.63%	3.77%	11.20	21.73%	4.73%	10.81	21.22%	3.84%	15.97	27.21%	9.69%
LSTM	SAEs	9.53	19.32%	3.90%	11.57	22.55%	5.08%	12.79	24.50%	5.74%	14.77	25.94%	7.90%
	LSTM	9.66	19.51%	4.02%	11.43	25.54%	5.00%	12.76	27.36%	5.70%	14.33	28.45%	7.44%
	GRU	9.53	24.13%	3.81%	11.75	22.58%	5.34%	13.08	24.57%	6.13%	14.95	25.99%	8.17%
GRU	SAEs	9.75	19.84%	4.08%	10.72	21.34%	4.10%	11.03	21.33%	4.09%	14.75	25.74%	8.00%
	LSTM	9.65	23.59%	3.95%	10.71	24.75%	4.20%	11.22	26.23%	4.31%	14.29	28.61%	7.50%
	GRU	9.87	19.92%	4.19%	11.00	21.52%	4.46%	11.51	21.85%	4.60%	14.91	25.85%	8.24%

By comparing the results of Table 2, the prediction method with GRU-based imputation can be safely recommended for traffic flow prediction under the DMR pattern.

Next, we investigate the performance of different deep learning based traffic flow prediction methods under the BDM model, as shown in Table 3. It is clear that the traffic flow prediction methods with deep learning based imputation outperform the mean imputation based prediction methods. Moreover, the traffic flow prediction methods with deep learning based imputation have a comparable performance. Table 4 evaluates the prediction accuracy of the traffic flow prediction methods under the MBDM model. Although the performance of MOL imputation method in higher missing rates is better than the deep learning

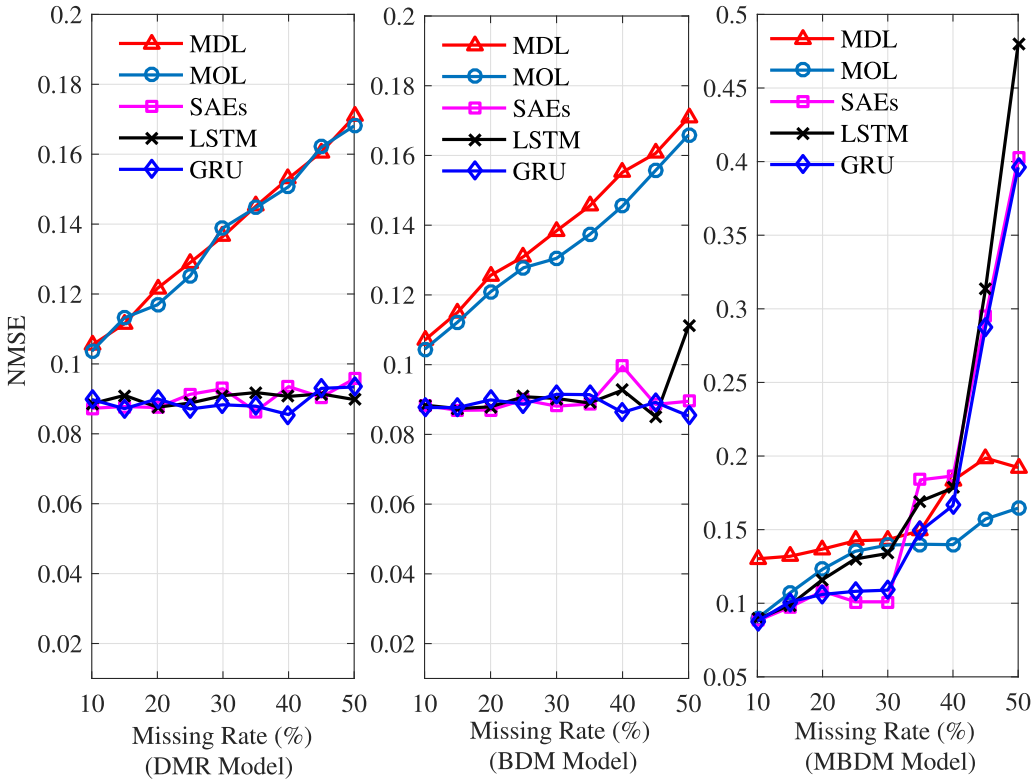


FIGURE 7. Comparison of NMSE for different methods based on different model.

based imputation methods, the prediction methods with deep learning based imputation still achieve a approximative performance with the MOL imputation based traffic flow prediction method.

To better understand the results of the experiments, the comparison of NMSE for different methods based on different models is shown in Fig. 7. On the DMR and BDM model, the deep learning based imputation methods has a noticeable improvement in the NMSE of traffic prediction. However, when the missing data rate is beyond 40 percent, it has a visible influence on the accuracy of traffic flow prediction on the MBDM model. Compared with the mean-based imputation methods, the deep learning based prediction methods are still recommended as the more efficient methods to predict the traffic flow with data missing due to their ability of extracting the internal features among training data. Specifically, SAEs represents the fully-connected network method without memory replay, and RNN optimizes the network by adding the loop structure. GRU and LSTM are both RNN-based methods, and GRU performs better in overcoming the slow response problem of the LSTM algorithm by capturing the long-term dependencies from training data. Moreover, we could find the similar performance and patterns of MAE and MRE results from Table 2 to Table 4.

V. CONCLUSION

In this paper, we conduct a comprehensive study to the prediction of traffic flow using an incomplete data set and via deep learning approaches. The DLIP model is proposed to solve the imputation and prediction of traffic flow. To capture the impact from missing data, we introduce three different patterns to model the missing data. Moreover, we also propose two types of methods for the imputation, i.e., the mean and deep learning based imputation methods, to increase the accuracy of traffic flow prediction. Based on the results of tests, the appropriate traffic flow prediction approaches under different data missing models have been thoroughly discussed and design guidelines for a practical implementations are also given.

Furthermore, the proposed DLIP model provides a proper option for solutions on prediction problems of wireless communication, data processing, and engineering with missing training data in future work. For instance, the efficient DLIP based prediction method can be implemented to solve the channel estimation problem under the situation losing enough measurement of training channel data in practice.

REFERENCES

[1] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 865–873, Apr. 2015.

[2] X. Zhou, X. Cai, Y. Bu, X. Zheng, J. Jin, T. H. Luan, and C. Li, "When road information meets data mining: Precision detection for heading and width of roads," *IEEE Access*, vol. 7, pp. 60399–60410, 2019.

[3] W. Huang, G. Song, H. Hong, and K. Xie, "Deep architecture for traffic flow prediction: Deep belief networks with multitask learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 5, pp. 2191–2201, Oct. 2014.

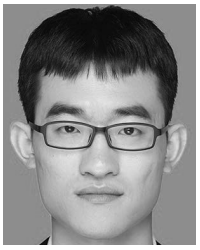
[4] J. Zhao, Q. Li, Y. Gong, and K. Zhang, "Computation offloading and resource allocation for cloud assisted mobile edge computing in vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 7944–7956, Aug. 2019.

[5] X. Sun, J. Zhao, X. Ma, and Q. Li, "Enhancing the user experience in vehicular edge computing networks: An adaptive resource allocation approach," *IEEE Access*, vol. 7, pp. 161074–161087, 2019.

[6] B. Ghosh, B. Basu, and M. O'Mahony, "Multivariate short-term traffic flow forecasting using time-series analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 2, pp. 246–254, Jun. 2009.
[7] S. Sun, C. Zhang, and G. Yu, "A Bayesian network approach to traffic flow forecasting," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 1, pp. 124–132, Mar. 2006.
[8] E. I. Vlahogianni, J. C. Golias, and M. G. Karlaftis, "Short-term traffic forecasting: Overview of objectives and methods," *Transp. Rev.*, vol. 24, no. 5, pp. 533–557, Feb. 2007.
[9] C. Li, Y. Fu, F. R. Yu, T. H. Luan, and Y. Zhang, "Vehicle position correction: A vehicular blockchain networks-based GPS error sharing framework," *IEEE Trans. Intell. Transp. Syst.*, early access, doi: 10.1109/TITS.2019.2961400.
[10] J. Zhao, S. Ni, and Y. Gong, "Peak-to-Average power ratio reduction of FBMC/OQAM signal using a joint optimization scheme," *IEEE Access*, vol. 5, pp. 15810–15819, May 2017.
[11] S. Ni, J. Zhao, H. H. Yang, and Y. Gong, "Enhancing downlink transmission in MIMO HetNet with wireless backhaul," *IEEE Trans. Veh. Technol.*, vol. 68, no. 7, pp. 6817–6832, Jul. 2019.
[12] G. E. A. P. A. Batista and M. C. Monard, "An analysis of four missing data treatment methods for supervised learning," *Appl. Artif. Intell.*, vol. 17, nos. 5–6, pp. 519–533, May 2003.
[13] T. Liu, B. Tian, Y. Ai, L. Li, D. Cao, and F.-Y. Wang, "Parallel reinforcement learning: A framework and case study," *IEEE/CAA J. Automatica Sinica*, vol. 5, no. 4, pp. 827–835, Jul. 2018.
[14] Y. Qin, C. Wei, X. Tang, N. Zhang, M. Dong, and C. Hu, "A novel nonlinear road profile classification approach for controllable suspension system: Simulation and experimental validation," *Mech. Syst. Signal Process.*, vol. 125, pp. 79–98, Jun. 2019.
[15] W. Liu, D. Wei, and F. Zhou, "Fault diagnosis based on deep learning subject to missing data," in *Proc. Chin. Control Decis. Conf. (CCDC)*, Shenyang, China, Jun. 2018, pp. 3972–3977.
[16] Y. Duan, Y. Lv, W. Kang, and Y. Zhao, "A deep learning based approach for traffic data imputation," in *Proc. 17th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Qingdao, China, Oct. 2014, pp. 912–917.
[17] J. Poulos and R. Valle, "Missing data imputation for supervised learning," *Appl. Artif. Intell.*, vol. 32, no. 2, pp. 186–196, Mar. 2018.
[18] Y. Chen, Y. Lv, and F.-Y. Wang, "Traffic flow imputation using parallel data and generative adversarial networks," *IEEE Trans. Intell. Transp. Syst.*, early access. [Online]. Available: https://ieeexplore.ieee.org/document/8699108.
[19] N. G. Polson and V. O. Sokolov, "Deep learning for short-term traffic flow prediction," *Transp. Res. C, Emerg. Technol.*, vol. 79, pp. 1–17, Jun. 2017.
[20] L. Qu, J. Hu, L. Li, and Y. Zhang, "PPCA-based missing data imputation for traffic flow volume: A systematical approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 3, pp. 512–522, Sep. 2009.
[21] L. N. N. Do, H. L. Vu, B. Q. Vo, Z. Liu, and D. Phung, "An effective spatial-temporal attention based neural network for traffic flow prediction," *Transp. Res. C, Emerg. Technol.*, vol. 108, pp. 12–28, Nov. 2019.
[22] L. Wang, Z. Zhang, and J. Chen, "Short-term electricity price forecasting with stacked denoising autoencoders," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 2673–2681, Jul. 2017.
[23] D. Jo, B. Yu, H. Jeon, and K. Sohn, "Image-to-image learning to predict traffic speeds by considering area-wide spatio-temporal dependencies," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1188–1197, Feb. 2019.
[24] X. Wang, R. Jiang, L. Li, Y. Lin, X. Zheng, and F.-Y. Wang, "Capturing car-following behaviors by deep learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 910–920, Mar. 2018.
[25] Y. Lin, X. Dai, L. Li, and F.-Y. Wang, "Pattern sensitive prediction of traffic flow based on generative adversarial framework," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 6, pp. 2395–2400, Jun. 2019.
[26] J. Wang, L. Zhang, Q. Guo, and Z. Yi, "Recurrent neural networks with auxiliary memory units," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1652–1661, May 2018.
[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[28] J. Mackenzie, J. F. Roddick, and R. Zito, "An evaluation of HTM and LSTM for short-term arterial traffic flow prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 5, pp. 1847–1857, May 2019.
[29] J. Zhao, Y. Gao, Y. Qu, H. Yin, Y. Liu, and H. Sun, "Travel time prediction: Based on gated recurrent unit method and data fusion," *IEEE Access*, vol. 6, pp. 70463–70472, Dec. 2018.
[30] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, Dec. 1976.
[31] M. S. Santos, R. C. Pereira, A. F. Costa, J. P. Soares, J. Santos, and P. H. Abreu, "Generating synthetic missing data: A review by missing mechanism," *IEEE Access*, vol. 7, pp. 11651–11667, Feb. 2019.
[32] X. Chen, Z. Wei, Z. Li, J. Liang, Y. Cai, and B. Zhang, "Ensemble correlation-based low-rank matrix completion with applications to traffic data imputation," *Knowl.-Based Syst.*, vol. 132, pp. 249–262, Sep. 2017.
[33] 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 Things J.*, vol. 6, no. 4, pp. 6618–6628, Aug. 2019.
[34] Caltrans. (2014). *Performance Measurement System (PeMS)*. [Online]. Available: http://pems.dot.ca.gov

JUNHUI ZHAO (Senior Member, IEEE) received the M.S. and Ph.D. degrees from Southeast University, Nanjing, China, in 1998 and 2004, respectively. From 1998 to 1999, he worked with the Nanjing Institute of Engineers, ZTE Corporation. Then, he worked as an Assistant Professor with the Faculty of Information Technology, Macao University of Science and Technology, in 2004, and continued there as an Associate Professor, until 2007. In 2008, he joined Beijing Jiaotong University, as an Associate Professor, where he is currently a Professor with the School of Electronics and Information Engineering. Since 2016, he has been with the School of Information Engineering, East China Jiaotong University. Meanwhile, he was also a short-term Visiting Scholar with Yonsei University, South Korea, in 2004, and a Visiting Scholar with Nanyang Technological University, Singapore, from 2013 to 2014. His current research interests include wireless and mobile communications and its related applications, which contain 5G mobile communication technology, high-speed railway communications, vehicle communication networks, wireless localization, and cognitive radios.

YIWEN NIE received the B.Eng. degree in computer science from East China Jiaotong University, Nanchang, China, in 2017. He is currently pursuing the Ph.D. degree with Beijing Jiaotong University, Beijing. His research interests include machine learning, vehicular communication, and UAV-aided communication.



SHANJIN NI received the Ph.D. degree from Beijing Jiaotong University, in 2019. He is with the National Computer Network Emergency Response Technical Team/Coordination Center of China (CNCERT/CC). His work focuses on the exploitation of massive multiple-input multiple-output (MIMO) communication and millimeter waves (mm-waves) for 5G HetNets. His research interests include pilot designs for multicell massive MIMO systems, pilot contamination reduction, and hybrid precoding for mm-wave massive MIMO.

XIAOKE SUN received the B.E. degree in electronic information science and technology from Xiangtan University, Hunan, China, in 2016. She is currently pursuing the Ph.D. degree in communication and information systems with Beijing Jiaotong University, Beijing, China. Her research interests include 5G vehicular networks, mobile edge computing, resource allocation, and stochastic optimization.