IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 15 March 2023, accepted 29 March 2023, date of publication 31 March 2023, date of current version 5 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3263720

RESEARCH ARTICLE

Heterogeneous Traffic Flow Detection Using CAV-Based Sensor With I-GAIN

BOWEN GONG[1], ZHIPENG XU[1], CIYUN LIN[1], AND DAYONG WU[2]
[1]Department of Traffic Information and Control Engineering, Jilin University, Changchun 130022, China
[2]Texas A&M Transportation Institute, Texas A&M University, College Station, TX 77843, USA

Corresponding author: Ciyun Lin (linciyun@jlu.edu.cn)

This work was supported in part by the Scientific Research Project of the Education Department of Jilin Province under Grant JJKH20221020KJ.

ABSTRACT This study proposes using connected automatic vehicles (CAVs) as traffic flow detectors to collect and exchange traffic flow data for heterogeneous traffic management and control. The proposed method includes the construction of a mathematical matrix to represent the status of the road section, the use of unsupervised machine learning to evaluate traffic data, and an improved generative adversarial imputation net (GAIN) to evaluate and impute missing traffic data. Next-generation simulation (NGSIM) data are used to verify the accuracy and robustness of the proposed method. One of the primary innovations of this study is the use of GAIN, a deep learning framework based on generative adversarial networks (GANs), to impute missing traffic data. GAIN has been shown to be more robust and stable when handling incomplete heterogeneous data than existing imputation methods. Additionally, this study contributes to the field by proposing the use of CAVs as sensors to detect mixed traffic flow, which could lead to more efficient and accurate traffic management and control. Experimental results demonstrate that the proposed method outperforms existing imputation methods, with a normalized root mean squared error and symmetric mean absolute percentage error of less than 0.2/0.3 and 0.08/0.13 in I-80 and Lankershim Boulevard, respectively. The findings of this study have important implications for the development and implementation of connected and automated vehicle technologies in the field of transportation.

INDEX TERMS CAV-based sensor, heterogeneous traffic flow, mixed traffic flow, improved generative adversarial imputation net, NGSIM.

I. INTRODUCTION

Multisensors are an essential part of connected automatic vehicles (CAVs) and can perceive and interacts with road conditions, traffic, and the driving environment, allowing CAVs to make trajectory planning and driving decisions. Therefore, the traffic network's safety, mobility, and efficiency can be markedly enhanced when CAVs replace conventional human-driven vehicles (HDVs) or have a relatively high market penetration rate [1]. However, this process will be gradual as CAVs displace HDVs. The phenomenon of traffic flow mixed with CAVs and HDVs is expected to last for at least the next 50 years [2].

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

Whether in the pure HDV environment or fully CAV traffic flow conditions, real-time traffic information is critical to traffic management and control to improve traffic efficiency and reduce traffic accidents. When the traffic flow is mixed with CAVs and HDVs, their heterogeneous performance and mutual interference will change the dynamics, safety, and mobility of traffic flow. This situation requires more comprehensive traffic perception and traffic state identification for traffic management and control to enhance traffic safety, improve road utilization, and reduce traffic congestion.

CAVs have sensors that perceive surrounding vehicles' location, speed, and direction for trajectory planning and driving operations to avoid collisions and improve safety. Therefore, CAVs as traffic flow detectors can collect and exchange traffic information for heterogeneous traffic management and control [3]. Scholars have focused on using the

data acquired by CAVs to estimate traffic status and travel time, and can include complementary traffic data for traffic management and control systems to improve control precision and efficiency [4], [5], [6], [7]. However, past studies have primarily focused on pure CAV traffic environments [8]. To our knowledge, using CAV-based sensors to estimate the mixed traffic flow status has rarely been addressed.

In this paper, we propose an improved generative adversarial imputation net algorithm (I-GAIN) to evaluate the traffic state for mixed traffic flow using CAV-based sensors [9]. First, a road is divided into grids, and a matrix is constructed to present the status of the road section. The CAV location is matched to the road section to identify the traffic data sensing status. Second, an unsupervised machine learning algorithm is used to evaluate the traffic data for road sections that are not CAVs. An improved generative adversarial imputation net is then proposed to enhance the precision of traffic data for road sections without CAVs. Finally, the Next Generation Simulation (NGSIM) dataset is used to validate the effectiveness and accuracy of the proposed method [10].

The primary contributions of this study are as follows:

(1) A mathematics matrix is proposed to present the traffic data collection status of the road section by mapping the CAVs into the matrix.

(2) An improved I-GAIN algorithm is proposed to impute the missing traffic data without CAVs on the road section or out of the CAV perception range.

(3) The accuracy and robustness of the proposed method are effective and perform better with NGSIM data for verification and comparison.

The remainder of this paper is organized as follows. Section II reviews related research on traffic sensing for mixed traffic flow. Section III presents the problem description. Section IV describes the proposed method. In Section V, the computational efficiency of the method is reported. Experiments and analyses are conducted to describe the performance of the method. Section VI concludes this paper with a summary of the contributions and limitations of the proposed model, as well as perspectives on future work.

II. LITERATURE REVIEW

To draw a clear distinction between this and previous studies, the literature on traffic information detection with camera-based sensors and light detection and range (LiDAR) sensors and CAV traffic perception is reviewed. The methods and algorithms for missing data imputation are also introduced in this section.

A. VISION-BASED TRAFFIC INFORMATION DETECTION

Recently, camera-based vision sensors have been widely used in traffic information detection due to their low cost and ability to provide rich perception information. Wei [11] used a histogram of oriented gradients (HOG) and Harr features to segment the region of interest and extract targets, solving the multivehicle detection problem in complex driving environments. Wang [12] developed a real-time target detection system using a field-programmable gate array board that converted color images to grayscale maps and extracted HOG features from maps of different sizes based on the HOG method. Xu [13] proposed an adversarial Faster-RCNN algorithm based on global averaging pooling to generate complex samples for better object detection models. However, the measurements of camera-based sensors can be adversely affected by changes in lighting and adverse weather conditions. Additionally, such sensors cannot directly obtain depth and location information, which affects the accuracy and robustness of traffic parameter precision [14].

B. LIDAR-BASED TRAFFIC INFORMATION DETECTION

LiDAR, a modern active visual sensor, has various advantages, such as anti-interference to external light changes, adaptability to complex environments, broad scanning coverage, and rich perception information [15], [16], [17]. Zhao [18] developed a systematic approach to detect and track pedestrians and vehicles using 16 laser LiDAR sensors, with an average accuracy of 95% in traffic detection, classification, and tracking. Lin [17] proposed a lane detection algorithm for low-density roadside LiDAR, which can aid in high-precision vehicle positioning in vehicle-to-infrastructure (V2I) cooperation applications within intelligent transportation systems. Liu [15] proposed a novel static background construction method that used the fast Fourier transform (FFT) to classify distant target points and noise points with sparse point clouds to expand the detection range of low-channel roadside LiDAR. Zhang [19] introduced an unsupervised clustering method for roadside LiDAR applications that relies on a region-growing algorithm coupled with component labeling and a revised merging process to maintain high accuracy while improving computation speed and oversegmentation. However, the processing of point cloud data for LiDAR sensors demands a lot of computational resources and is time-consuming, which can hinder its real-time application in engineering.

C. CAV WITH TRAFFIC INFORMATION PERCEPTION

Researchers have been motivated to use CAV as a ''mobile sensor'' to collect traffic information due to their powerful sensory ability. Zheng [20] predicted traffic volumes at an intersection by extracting GPS data from CAVs and considering a maximum likelihood problem. Li [21] developed a cooperative perception framework using data collected by CAVs to predict the traffic state of a platoon of CAVs. Wei [22] proposed a three-step evolution strategy of the CAV perception mode to enhance urban transportation efficiency. Day [23] optimized signal coordination using CAV data in a low penetration rate environment. Li [24] used CAVs as an alternative data source for freeway traffic management, developing an interval type 2 fuzzy logic-based variable speed limit (VSL) system for mixed traffic to manage inherent uncertainty. With the development of vehicle-to-everything (V2X) and self-driving technology, more studies

use CAVs as detectors to collect high-resolution microlevel traffic data for traffic management and control systems, including vehicle-infrastructure cooperation systems. However, when the market penetration rate of CAVs is low, their perception range may not cover the entire road, leading to missing data. Therefore, future research should thus consider this limitation and focus on developing methods to improve data coverage in low CAV penetration rate scenarios.

D. IMPUTING MISSING DATA VALUES

Because missing data are ubiquitous in many domains and missing data imputation can help improve measurement accuracy and model performance, many data imputation methods have been proposed. Researchers have used single values to fill in the missing values to create many imputation methods, including mean imputation [25], hot deck imputation [26], cold deck imputation [27], and regression imputation [28]. However, using a single value to impute missing values, will produce an imputed dataset that has a certain degree of uncertainty, and the distribution of the imputed data will distort the distribution of the original sample, which will lead to bias in the data analysis results. To compensate for the shortcomings of the single imputation method, researchers have used multiple imputation methods (MIs) to impute missing data, and MIs include regression prediction, multiple regression imputation, propensity score, logistic regression, discriminant analysis, and Markov Chain Monte Carlo (MCMC) models [29], [30], [31]. In addition, with the development of deep learning, several researchers have developed deep learning frameworks based on autoencoder (AE) and generative adversarial networks (GAN) to impute missing data, which can obtain better robustness and relative stability in handling incomplete heterogeneous missing data [9], [32], [33], [34]. Zhang [35] proposed a self-attention generative adversarial imputation net that combines a self-attention mechanism, an autoencoder, and a generative adversarial network. The introduction of the self-attention mechanism can help their model effectively capture correlations between spatially distributed sensors at different time points. Wang [36] proposed a novel Generative Adversarial Guider Imputation Network (GAGIN) based on generative adversarial network (GAN) for unsupervised imputation, which is composed of a Global-Impute-Net (GIN), a Local-Impute-Net (LIN) and an Impute Guider Model (IGM) to solve two problems: the local homogenous regions and the reason for the imputed data. Yuan [37] proposed a novel spatiotemporal GAN model for traffic data imputation (STGAN) to efficiently impute traffic data.

Although GAN has many advantages in data imputing, its disadvantage is also more important: the generated data may have bias, which will lead to poor quality of the padded data. Therefore, the goal of this paper is to improve GAIN to reduce the error of padding data and thus increase the accuracy of the CAV perception algorithm.

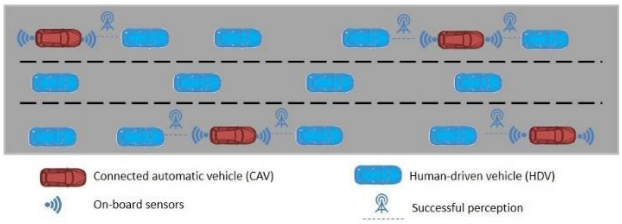

FIGURE 1. Traffic flows are mixed with HDVs and CAVs.

III. PROBLEM FORMULATION

Fig. 1 shows an example of mixed traffic flows with HDVs and CAVs, where HDVs are completely driven by the driver without perception capabilities, and CAVs are automatically driven with advanced assist driver systems based on onboard sensors and can exchange traffic data with the roadside unit. Each CAV can obtain traffic information within its perception range, while vehicles outside the perception range of the CAVs are not sensed.

In a mixed traffic flow, the perception capability of CAVs is limited. When there are no roadside LiDAR or other sensing devices in the road network but ''mobile sensor'' CAVs are present, the information collect by all vehicles on the road network may not be sensed by CAVs if the market penetration rate of CAVs is low. For example, a vehicle that is out of the range of any CAVs will not be sensed by the CAVs, and all information about them will be lost. This issue will affect mixed traffic flow management and control with regard to traffic safety and efficiency [38]. Therefore, this study primarily investigates how to accurately obtain traffic information in the road network without changing the CAV market penetration rate. To restrict influence factors, we assume the following:

(1) Each CAV is equipped with the same sensors, all of which have the same perception capabilities and are unaffected by environmental factors such as light change.

(2) CAVs are distributed randomly in the mixed traffic flow.

The notations in this problem are as follows.

IV. METHODS

The process of data imputation for mixed traffic flow includes two processes: modeling traffic state models with mathematics matrices, and traffic data imputation for HDVs not sensed by CAVs.

When modeling traffic states with mathematics matrices, we first input the data that are obtained directly by CAVs. Through the modeling traffic states with the matrices process, the data will be transformed into the data matrix (Da). Then, we perform the first imputation for Da to obtain the estimated matrix (E). Finally, E with the smallest error is selected from the first imputation for the second imputation. The imputed matrix (Im) and the error are then output. The error represents the gap between the imputed data and the real data, and is used as an assessment measure of imputation methods. An overview of the model is shown in Fig. 2.

TABLE 1. Symbols used in this paper and their interpretations.

Symbol	Interpretations
$\|\cdot\|$	Counting measure for the countable sets;
l	Index of a certain lane;
L	The set of all lane indices;
Da	Data matrix;
E	Estimated matrix;
Im	Imputed matrix;
R	Road matrix;
O	Original matrix;
Ma	Mask matrix;
m	A certain longitudinal location along the road;
M_l	Set of all longitudinal locations on lane;
s	Index of a certain longitudinal road segment;
S	Set of all indices s;
M_{ls}	Set of all longitudinal locations in road segment s and lane l;
i	Index of a certain vehicle;
I	Set of all vehicle indices i;
D^i	Detection area of a CAV;
I^A	Set of all CAVs indices;
D^{I^A}	Detection area D of all CAVs;
Q	Original matrix of vehicles information sensed by CAVs on a road;
Z	Random matrix;
H	Hint matrix;
$\widehat{Ma}$	Estimated mask matrix;

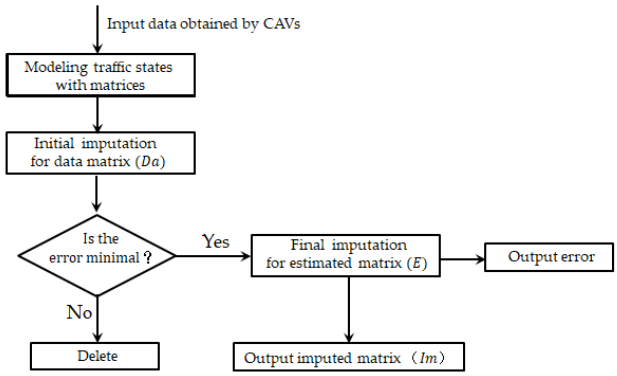

FIGURE 2. Flowchart of the proposed model.

A. SENSOR ERROR PREPROCESSING

1) SENSOR ERROR SOURCES

In this article, a car camera is used, and its accuracy varies with its price. This camera may have the following sources of error:

(1) Lighting conditions: the camera may shoot differently under different lighting conditions, which may affect the accuracy of the data.

(2) Position offset: Due to the different installation locations, the camera may have a position offset, resulting in errors in data collection.

(3) Low signal-to-noise ratio: When the signal-to-noise ratio is low, the camera may not be able to correctly identify and capture the target, leading to errors in data acquisition.

(4) Data loss: The camera may suffer from data loss, which also leads to errors in data acquisition.

2) ERROR ANALYSIS AND CORRECTION

To evaluate the effect of perception error on the data filling effect, we analyzed the relationship between sensor perception error and data filling error. We used the following mathematical model:

$$y = x + \epsilon \tag{1}$$

where x is the true traffic flow data, y is the sensed traffic flow data, and ϵ is the sensing error. We used the mean squared error (MSE) to measure the magnitude of the sensing error:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i - \epsilon_i)^2 \tag{2}$$

where n is the sample size; x_i is the true traffic flow data; y_i is the perceived traffic flow data; and $\in_i$ is the perception error.

To investigate the effect of perception errors on data imputation, we introduced several correction strategies for perception errors. For each flow data point, we assumed that its perception error followed a Gaussian distribution:

$$N\left(\mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{3}$$

Estimated these parameters by calculating the mean and variance of all the data in the current time window. Then, we used this distribution to correct the perception errors and obtain more accurate flow data.

Specifically, we assume a missing traffic data point $\widehat{x_i}$ that has a perceived value of $\widehat{y_i}$; then, we can calculate its true value as:

$$\widehat{x_i} = \widehat{y_i} - \widehat{\epsilon_i} \tag{4}$$

where $\widehat{\epsilon_i}$ is the perception error corrected by the Gaussian distribution, with an expected value and variance of, respectively:

$$\widehat{\mu_i} = \frac{1}{T}\sum_{t=1}^{T}(y_i - x_i)_t \tag{5}$$

$$\widehat{\sigma_i}^2 = \frac{1}{T-1}\sum_{t=1}^{T}((y_i - x_i)_t - \widehat{\mu_i})^2 \tag{6}$$

where T is the number of data points in the current time window. Therefore, we can correct the perception error using the following formula:

$$\widehat{\epsilon_i} = \widehat{\mu_i} + \widehat{\sigma_i} z_i \tag{7}$$

where z_i is a random variable that follows a standard normal distribution $N(0,1)$.

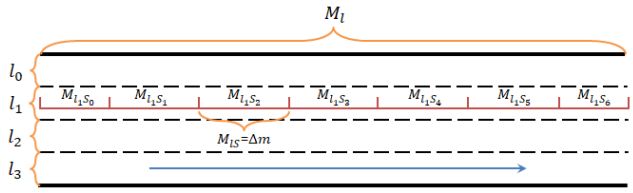

FIGURE 3. Schematic diagram of the road segment.

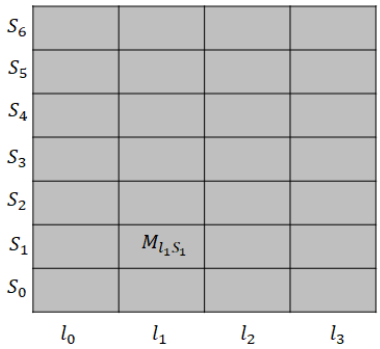

FIGURE 4. Overview of the road matrix.

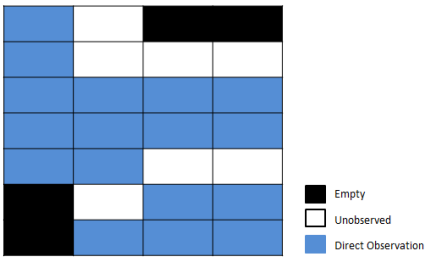

FIGURE 5. Overview of the original matrix.

B. MODELING TRAFFIC OF THE PROPOSED MODEL

We assume that there is a highway with L lanes and the lane number set is $l \in \{0, 1, 2, \ldots, L-1\}$. Each lane is regarded as a one-dimensional straight line, and the length of lane l is M_l. Each lane has the same length in this paper. M_l is divided into road segments, and the length of each road segment is S. The length of the road segment primarily depends on the length of the vehicle and the minimum spacing between vehicles. The road segments are numbered $\{0, 1, 2, \ldots, S-1\}$. Therefore, the road segment can be presented as M_{lS}, the subscript l means that the road segment in lane l, and the subscript S means that the road segment is the S^{th} segment. The location of M_{lS} is $[s\Delta m, (s+1)\Delta m]$, $\Delta m = M_l/S$, as shown in Fig. 3.

According to Fig. 3, we can describe the entire road as a matrix, which is called the road matrix (R). The columns and rows of the matrix represent the l and S of the road, respectively. As shown in Fig. 4, $M_{l_1S_1}$ is the region on lane l_1 and road section S_1.

After constructing R, we accurately fill in the traffic information obtained by CAVs directly according to the location of vehicles into R, which is the original matrix (O).

1	0		
1	0	0	0
1	1	1	1
1	1	1	1
1	1	0	0
	0	1	1
	1	1	1

FIGURE 6. Overview of the mask matrix.

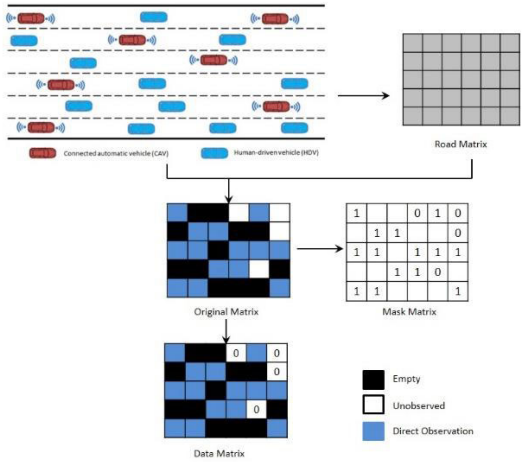

FIGURE 7. Modeling traffic states with matrices.

As shown in Fig. 5, the black cell of the matrix means that there is no vehicle on this road segment; the white cell means that the vehicle on this road segment cannot be directly sensed by CAVs; the blue cell means that the vehicle on this road segment can be directly sensed by CAVs; and no data exist in the black and white cells. Thus, there are three types of road segments. If the segments that cannot be directly sensed by CAVs and the segments without vehicles cannot be accurately identified, the segments without vehicles will also be imputed, which will lead to inaccurate experimental results. To mitigate this issue, we obtain a mask matrix (Ma) from O to solve this problem. Ma also plays an important role in the subsequent imputation.

As shown in Fig. 6, each cell of Ma equals of $[0, 1]^d$ or none. When Ma(l, S) is 1, M_{lS} can be directly sensed by CAVs. When Ma(l, S) is 0, M_{lS} cannot be directly sensed by CAVs. When Ma(l, S) is zero, there is no vehicle on M_{lS}. The relevant formula is:

$$Ma(l,S) = \begin{cases} 1 & \text{if } M_{lS} \in D^{I^A} \\ 0 & \text{if } M_{lS} \notin D^{I^A} \text{ and } \exists\, i \text{ on } M_{lS} \\ \text{none} & \text{else} \end{cases} \quad (8)$$

where I^A is the set of all CAVs and D^{I^A} is the total perception range of all CAVs on the road.

We assumed that onboard sensors are placed at the front and rear of each CAV, and that the perception range is

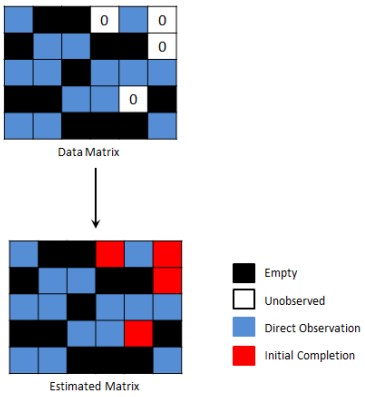

FIGURE 8. Initial imputation.

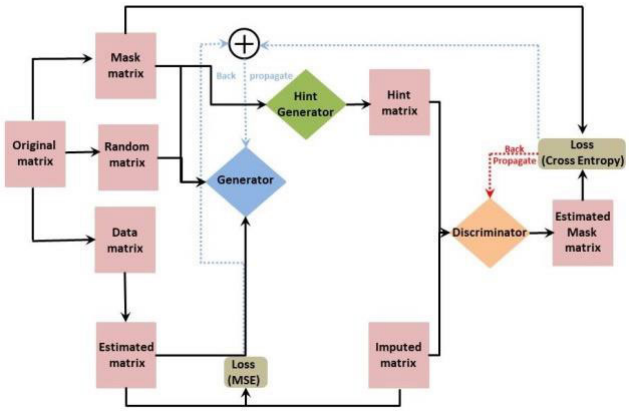

FIGURE 9. Flowchart of the I-GAIN algorithm.

20-60 m [39], as shown in Fig. 7. Therefore, CAVs can only obtain information about the vehicles in front of and behind them in their lane, not information about other vehicles. Because the road has 6 lanes, the matrix has 6 columns. The matrix is divided into 5 rows based on the road length, vehicle length, and minimum spacing between vehicles.

A cell with missing data in O (a white cell) is imputed to be equal to 0 to obtain Da. Improving the perception capability of CAVs can be transformed into a matrix imputation problem by modeling traffic states with matrices.

C. DATA IMPUTATION FOR HDVs

1) INITIAL IMPUTATION FOR DATA MATRIX

As shown in Fig. 8, E can be obtained from Da by the initial imputation. The red cell means that the data in this cell are imputed with the initial imputation.

The unsupervised mechanical algorithm simple fill (SF), k-nearest neighbor imputation (KNN) [42], iterative imputer (II) [43], and matrix factorization (MF) [44] were used to improve the accuracy of the initial imputation. All four initial imputation algorithms were used to ensure that the data of the nonmissing part of the matrix were unchanged, and only the missing part of the matrix could be imputed.

First, SF takes the average of each column of the matrix to impute. For the KNN algorithm, the mean squared difference of the features of the observed data in both rows is used to weigh the samples, and then, the weighted results are used to fill the eigenvalues. The K with the best imputation effect is selected using the principle of "the closer the better" to impute the missing values of the target features with the distribution of other features, which will be more reliable than imputing directly with the mean and median. The steps of the KNN algorithm are as follows.

(1) Input O and find the K nearest samples closest to the missing data using Euclidean Distance in the matrix. Euclidean Distance d_{ls} is:

$$d_{ls} = \sqrt{w \times p} \tag{9}$$

$$w = \frac{N}{n} \tag{10}$$

where w is the weight of the sample, p is the squared distance from the present coordinates, N is the total number of coordinates and n is the number of present coordinates.

(2) Missing values are imputed using the mean of the nonempty values of the corresponding positions of the K nearest neighbors.

(3) Output the imputed value and its location.

Algorithm II imputes missing values by modeling each feature with a missing value as a function of other features in a cyclic manner. This strategy models each feature with a missing value as a function of other features. The steps of algorithm II are as follows.

(1) Input O and proceed in an iterative loop.

(2) At each step, one feature column is specified as the output y, and the other feature columns are treated as the input X.

(3) A regressor fits (X, y) to a given y. The regressor is then used to predict the missing value of y.

(4) The max_iter imputation wheel is repeated and outputs the result of the last round of imputation.

MF decomposes the incomplete matrix directly into low-rank "U" and "V". Then, the gradient descent method is used to solve the matrix factorization: it can reduce the computation amount and can solve the sparse behavior matrix problem caused by the number of users and too many items. MF is:

$$J = \min \left\| O - \overline{\overline{O}} \right\| = \left\| O - UV^T \right\|^2 = \sum_{l.s, o_{ls} \neq nan} \left(o_{ls} - \sum_{j=1}^{k} v_{lj} u_{sj} \right)^2 \tag{11}$$

where $\overline{\overline{O}}$ is the approximate matrix of O, and v_{lj} and u_{sj} are the elements of U and V respectively, which is what we want to determine.

These four algorithms perform the initial imputation on Da in succession. The algorithm with the minimum error value is selected to prepare for the final imputation.

2) FINAL IMPUTATION FOR ESTIMATED MATRIX

Because the accuracy of the initial imputation is low, the final imputation can be performed to improve the accuracy. In this study, GAIN is used as the final imputation algorithm and is an unsupervised imputation method that can be applied to any type of data. GAIN also does not require complete data for training and can obtain higher accuracy. After the initial imputation, the algorithm with the minimum error value is selected from the initial imputation, and then, GAIN is used for the final imputation, which is called I-GAIN in this paper. The flowchart of the I-GAIN algorithm is shown in Fig. 9.

The random matrix (Z) is the matrix that simulates the random noise, and the position of the missing data is recorded by Ma. E, Z, and Ma will be used as the input of the generative network, and Im as the output of the generator. Im and the hint matrix (H) are used to represent the location and randomness of missing data as the input of the discriminative network. The output of the discriminative network is the estimated mask matrix ($\widehat{\mathrm{Ma}}$). The value of each cell represents the authenticity of the data at that location. The range of values is from 0 to 1, with the truest value being 1 and the least false value being 0. The loss functions are the reconstructed error calculated by E and Im, as well as the cross entropy by $\widehat{\mathrm{Ma}}$ and Ma.

The generative network and the discriminative network are updated iteratively by the back propagation method until the loss converges. In this case, the discriminative network and the generative network are both strong, and the generative network can impute the missing data to be closer to the real data perfectly. This process is a game process. First, the generative network and the discriminative network are weak. To achieve the game victory, the generative network keeps optimizing itself to make the generated data increasingly realistic so that the discriminative network cannot identify the fake data. The discriminative network keeps optimizing itself to improve the discriminative ability, which can correctly distinguish real and imputed data. Eventually, this process will reach a balanced state, in which the generative network can generate more realistic data and the discriminative network has stronger discriminative power.

Q is the original matrix of vehicle information sensed by CAVs on a road. Vehicle information includes speed, acceleration, traffic flow, density, etc. $q = (q_1, q_2, q_3 \ldots\ldots q_d)$ is the Q vector corresponding to an observation record of vehicle information, and $ma = (ma_1, ma_2, ma_3 \ldots\ldots ma_d)$ is the mask vector corresponding to an observer. The formula of q is:

$$q = \begin{cases} q_i & \text{if } ma_i = 1 \\ 0 & \text{if } ma_i = 0 \end{cases} \tag{12}$$

where q is the vector of the data matrix $\ddot{\mathrm{Q}}$; $\tilde{\mathrm{Q}}$ is the estimated matrix obtained by the initial imputation. The following equations describe these parameters:

$$\tilde{\mathrm{Q}} = \Psi(\ddot{\mathrm{Q}}, \mathrm{Ma}) \tag{13}$$

where ψ is the algorithm for initial imputation, and:

$$\tilde{q} = \begin{cases} q_i & \text{if } ma_i = 1 \\ \tilde{q}_i & \text{if } ma_i = 0 \end{cases} \tag{14}$$

where $\tilde{q}_i$ is the vehicle information imputed by the initial imputation process.

For the final imputation, each cell in Z is a random number from the distribution U (0, 1).

We assume that B is a random variable and $b = (b_1, b_2, b_3 \ldots\ldots b_d) \in [0, 1]^d$ is the corresponding B vector:

$$b_j = \begin{cases} 1 & \text{if } j \neq k \\ 0 & \text{if } j = k \end{cases} \tag{15}$$

where b_j is the jth value of vector b and k is the first sampling and $k \in \{1, \ldots, d\}$.

H is:

$$H = b \cdot \mathrm{Ma} + 0.5(1 - b) \tag{16}$$

and the generative network G is:

$$\overline{\mathrm{Q}} = G(\tilde{\mathrm{Q}}, \mathrm{Ma}, (1 - \mathrm{Ma}) \cdot Z) \tag{17}$$

The generative network G takes Ma, Z and $\tilde{\mathrm{Q}}$ as inputs and outputs the imputed matrix $\overline{\mathrm{Q}}$.

The discriminative network D is:

$$\widehat{\mathrm{Ma}} = D(\overline{Q}, H) \tag{18}$$

and takes $\overline{\mathrm{Q}}$ and H as inputs and outputs of the estimated mask matrix $\widehat{\mathrm{Ma}}$.

The pseudocode of the I-GAIN algorithm is shown in Table 2.

V. EXPERIMENTS AND EVALUATION

All experiments were performed in Python 3.6.12 on a computer equipped with an Intel (R) Xeon(R) CPU E5-2450 0 @ 2.10 GHz and 16.0 GB of RAM. The NGSIM data were used to conduct numerical experiments to verify the improved GAIN.

A. DATA AND EXPERIMENT SETUPS

In this section, NGSIM data were used to verify the proposed method. NGSIM data are high-resolution vehicle trajectory data on different roads [10]. The experiment was performed on the I-80 highway and Lankershim Boulevard, an urban road. In this paper, only datasets related to vehicle speed on I-80 and Lankershim Boulevard were used, but the proposed method is not only limited to speed; it is possible to use any information about vehicles, including acceleration, traffic flow, and density. The overview of I-80 and Lankershim Boulevard is shown in Fig. 10 and Fig. 11, respectively.

First, the I-80 highway is taken as R. Each cell of R has no more than one car, and the car length is limited to 6-7 m, with a minimum spacing of 7 m on the I-80. Thus, every cell in R represents a 15 m road segment. Due to the total length of 503 m, there are 34 segments per lane and 34 rows in R.

TABLE 2. Pseudocode of I-GAIN.

Algorithm 1 Pseudocode of I-GAIN

Input: An original matrix $\mathbf{Q}$ with missing data, the initial imputation algorithm $\boldsymbol{\psi}$, the number of iterations $\mathbf{T}$, a random matrix $\mathbf{Z}$, a random variable vector $\mathbf{b}$

Output: The imputed matrix $\overline{\mathbf{Q}}$

Step1: where $\ddot{\mathbf{Q}}$ is from $\mathbf{Q}$;//Calculate the data matrix $\ddot{\mathbf{Q}}$

Step2: where $\mathbf{Ma}$ is from $\mathbf{Q}$;//Calculate the mask matrix $\mathbf{Ma}$

Step3: $\widetilde{\mathbf{Q}} = \boldsymbol{\psi}(\ddot{\mathbf{Q}}, \mathbf{Ma})$;// Imputation $\ddot{\mathbf{Q}}$ with the initial imputation algorithm to obtain the estimated matrix$\widetilde{\mathbf{Q}}$

Step4: for(t=1 to $\mathbf{T}$) do

(1)Discriminator optimization

Draw x samples from the dataset$[(\widetilde{\mathbf{Q}}(\mathbf{j}), \mathbf{Ma}(\mathbf{j})]_{\mathbf{j=1}}^{\mathbf{x}}$

Draw x i.i.d. samples, $[\mathbf{z}(\mathbf{j})]_{\mathbf{j=1}}^{\mathbf{x}}$, of$\mathbf{Z}$

Draw x i.i.d. samples, $[\mathbf{b}(\mathbf{j})]_{\mathbf{j=1}}^{\mathbf{x}}$, of b

for j=1,…,x do

$\widetilde{\mathbf{Q}}(\mathbf{j}) \leftarrow \mathbf{G}(\widetilde{\mathbf{Q}}(\mathbf{j}), \mathbf{Ma}(\mathbf{j}), \mathbf{Z}(\mathbf{j}))$

$\widehat{\boldsymbol{Q}}(\mathbf{j}) \leftarrow \mathbf{Ma}(\mathbf{j}) \odot \widetilde{\mathbf{Q}}(\mathbf{j}) + (\mathbf{1} - \mathbf{Ma}(\mathbf{j})) \odot \overline{\mathbf{Q}}(\mathbf{j})$

$\mathbf{h}(\mathbf{j}) \leftarrow \mathbf{b}(\mathbf{j}) \odot \mathbf{Ma}(\mathbf{j}) + \mathbf{0.5}(\mathbf{1} - \mathbf{b}(\mathbf{j}))$

end for

Update D using stochastic gradient descent (SGD)

$$\nabla_{\mathbf{D}} - \sum_{\mathbf{j=1}}^{\mathbf{x}} \mathcal{L}_{\mathbf{D}}(\mathbf{Ma}(\mathbf{j}), \mathbf{D}(\widehat{\mathbf{Q}}(\mathbf{j}), \mathbf{h}(\mathbf{j}), \mathbf{b}(\mathbf{j}))$$

(2)Generator optimization

Draw x samples from the dataset$[(\widetilde{\mathbf{Q}}(\mathbf{j}), \mathbf{Ma}(\mathbf{j})]_{\mathbf{j=1}}^{\mathbf{x}}$

Draw x i.i.d. samples, $[\mathbf{z}(\mathbf{j})]_{\mathbf{j=1}}^{\mathbf{x}}$, of Z

Draw x i.i.d. samples, $[\mathbf{b}(\mathbf{j})]_{\mathbf{j=1}}^{\mathbf{x}}$, of b

for j=1,…,x do

$\mathbf{h}(\mathbf{j}) \leftarrow \mathbf{b}(\mathbf{j}) \odot \mathbf{Ma}(\mathbf{j}) + \mathbf{0.5}(\mathbf{1} - \mathbf{b}(\mathbf{j}))$

end for

Update G using SGD (for fixed D)

$$\nabla_{\mathbf{G}} \sum_{\mathbf{j=1}}^{\mathbf{x}} \mathcal{L}_{\mathbf{G}}(\mathbf{Ma}(\mathbf{j}), \widehat{\mathbf{Ma}}(\mathbf{j}), \mathbf{b}(\mathbf{j})) + \boldsymbol{\alpha}\mathcal{L}_{\mathbf{Ma}}(\mathbf{Q}(\mathbf{j}), \widetilde{\mathbf{Q}}(\mathbf{j}))$$

end for

$\overline{\mathbf{Q}} = \mathbf{G}(\widetilde{\mathbf{Q}}, \mathbf{Ma}, (\mathbf{1} - \mathbf{Ma}) \cdot \mathbf{Z})$

Step5:return$\overline{\mathbf{Q}}$

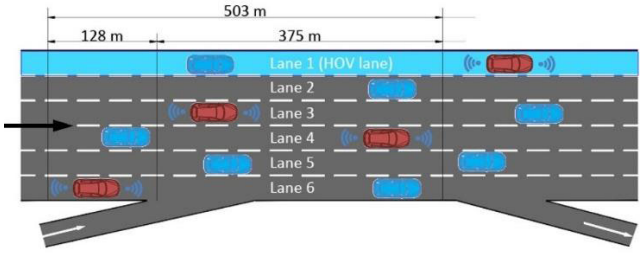

FIGURE 10. The overview of I-80.

because there are 6 lanes, R has 6 columns, and I-80 can be regarded as a 6-by-34 R.

The setup is the same on Lankershim Boulevard. Because Lankershim Avenue is an urban road and there are two signalized intersections on the road, the vehicle speed on the road will be lower, and the minimum spacing between vehicles will be smaller. The minimum spacing between vehicles is 1 m. According to the length of Lankershim Boulevard, each cell in R represents a 7 m road segment. In this experiment, there are 77 segments per lane; thus, there are 77 rows in R. Because there are 4 lanes, R has 4 columns, and Lankershim Boulevard can be regarded as a 4-by-77 R.

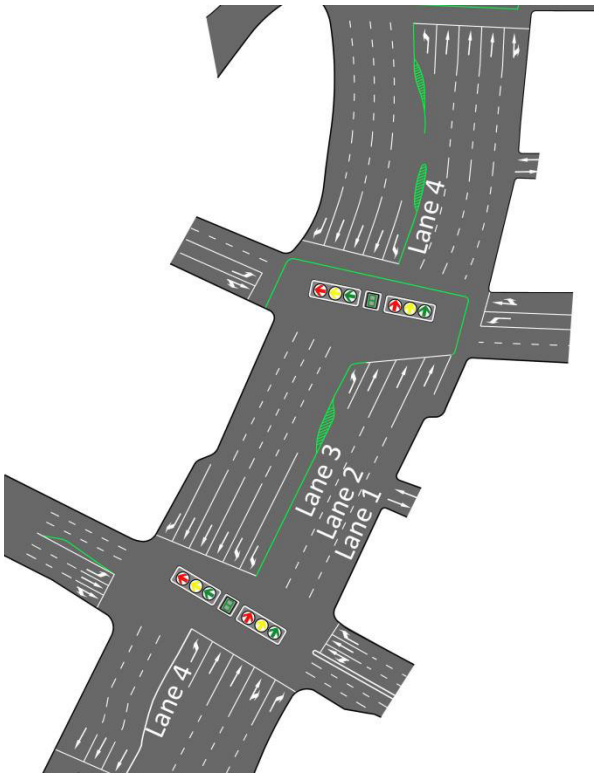

FIGURE 11. Overview of Lankershim Boulevard.

In the experiments, CAVs were randomly distributed on the road due to the limited perception range of CAVs. When the CAV market penetration rate is low, the vehicle information on the road cannot be completely obtained. The normalized root mean square error (NRMSE) and symmetric mean absolute percentage error (SMAPE1, SMAPE2) are the formulae used to determine the error values and are, respectively:

$$\mathrm{NRMSE}\left(v, \hat{v}\right) = \sqrt{\frac{\sum_{\mu \epsilon M} \left(v_\mu - \hat{v}_\mu\right)^2}{\sum_{\mu \epsilon M} v_\mu^2}} \tag{19}$$

$$\mathrm{SMAPE1}\left(v, \hat{v}\right) = \frac{1}{|M|} \sum_{\mu \epsilon M} \frac{|v_\mu - \hat{v}_\mu|}{v_\mu + \hat{v}_\mu} \tag{20}$$

$$\mathrm{SMAPE2}\left(v, \hat{v}\right) = \frac{\sum_{\mu \epsilon M} |v_\mu - \hat{v}_\mu|}{\sum_{\mu \epsilon M} \left(v_\mu + \hat{v}_\mu\right)} \tag{21}$$

where v is the true speed vector; $\hat{\mathrm{v}}$ is the imputed speed vector; μ is the index of the vector; and M is the set of indices in v and $\hat{\mathrm{v}}$. The output of the proposed method in this paper is a matrix, which must only be expanded into vectors before comparison.

In the baseline setting, the CAV market penetration rate is 20%; the CAV-sensed vehicle speed is not disturbed by noise; $|\mathrm{S}| = 15$ and $|\mathrm{L}| = 6$ on I-80; $|\mathrm{S}| = 7$ and $|\mathrm{L}| = 4$ on Lankershim Boulevard; and the detection range of the onboard sensor is 30 m.

TABLE 3. Initial imputation accuracy with the basic setting on I-80.

	SF	KNN	II	MF
NRMSE	0.2680	0.4042	0.2247	0.2208
SMAPE1	0.1219	0.2068	0.0914	0.0892
SMAPE2	0.1204	0.1585	0.0921	0.0875

TABLE 4. Initial imputation accuracy with the basic setting on Lankershim Boulevard.

	SF	KNN	II	MF
NRMSE	0.3938	0.7046	0.4024	0.3604
SMAPE1	0.1879	0.5279	0.1821	0.1680
SMAPE2	0.1405	0.3549	0.1379	0.1307

TABLE 5. Final imputation accuracy with the basic setting.

Location	NRMSE	SMAPE1	SMAPE2
I-80	0.1941	0.0785	0.0768
Lankershim Boulevard	0.2857	0.1232	0.1215

TABLE 6. Traffic sensing accuracy of different algorithms In I-80.

	MF	GAIN	SFG	KNNG	IIG	I-GAIN
NRMSE	0.2208	0.2406	0.2321	0.3016	0.2040	0.1941
SMAPE1	0.0892	0.1093	0.1056	0.1334	0.0892	0.0785
SMAPE2	0.0875	0.1047	0.1032	0.1312	0.0875	0.0768

B. ALGORITHM COMPARISON

In this paper, the baseline settings were selected to run the imputation methods. The total time of the initial imputation was 23 min, and the accuracy of each imputation method was the average value of each imputation result. Results are shown in TABLE 3 and TABLE 4.

TABLE 3 and TABLE 4 show that the accuracy of MF is the highest of the tested methods for both I-80 and Lankershim Boulevard. Then, based on MF, I-GAIN is used for the final imputation, and results are shown in TABLE 5.

Table 5 shows the final imputation results, which show a strong improvement in accuracy for both locations compared to the initial imputation. On I-80, the NRMSE decreased by 19.41%, and SMAPE1 and SMAPE2 decreased by 7.85% and 7.68%, respectively. On Lankershim Boulevard, the NRMSE decreased by 28.57%, and SMAPE1 and SMAPE2 decreased by 12.32% and 12.15%, respectively.

While MF achieved the best initial imputation performance, the final results achieved using I-GAIN markedly improved the accuracy for both locations. These findings suggest that I-GAIN can effectively refine imputation results obtained using base methods, such as MF. However, more experiments will be needed to determine the performance of I-GAIN compared to other imputation methods, as well as its suitability for different types of traffic data.

Therefore, MF, GAIN, SF-GAIN(SFG), KNN-GAIN (KNNG), and II-GAIN (IIG) were used to verify the accuracy and effectiveness of the proposed method with the same settings. Results are shown in TABLE 6 and TABLE 7.

TABLE 7. Traffic sensing accuracy of different algorithms in Lankershim Boulevard.

	MF	GAIN	SFG	KNNG	IIG	I-GAIN
NRMSE	0.3604	0.3589	0.3125	0.4232	0.3287	0.2857
SMAPE1	0.1680	0.1657	0.1375	0.1923	0.1428	0.1232
SMAPE2	0.1307	0.1312	0.1245	0.1425	0.1268	0.1215

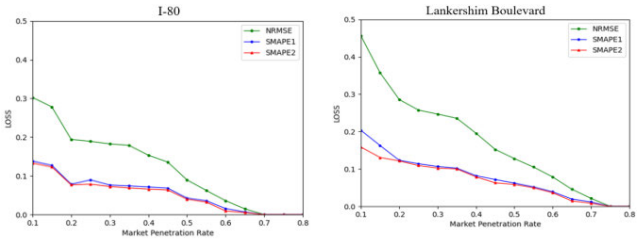

FIGURE 12. Traffic sensing accuracy under different CAV market penetration rates. (a) Accuracy on I-80. (b) Accuracy on Lankershim Boulevard.

The experiment evaluated the performance of six different algorithms for traffic sensing on I-80 and Lankershim Boulevard. The evaluation was based on the normalized root-mean-square error (NRMSE) and two symmetric mean absolute percentage error (SMAPE) metrics.

Tables 6 and 7 show the accuracy of the six algorithms on I-80 and Lankershim Boulevard, respectively. I-GAIN is shown to outperform all other algorithms in terms of accuracy, with NRMSE of 19.41% and 28.57%, and SMAPE1 and SMAPE2 of 7.85% and 7.68%, respectively, on I-80; and an NRMSE of 0.2857 and SMAPE1 and SMAPE2 of 12.32% and 12.15%, respectively, on Lankershim Boulevard.

Comparing Tables 5, 6, and 7 shows that I-GAIN improves the accuracy of traffic sensing by 34.15% and 20.96% on I-80 and Lankershim Boulevard, respectively, when compared with the best-performing algorithm among the other five methods.

Overall, experimental results demonstrate that the proposed I-GAIN algorithm performs better than existing methods for traffic sensing and t can effectively improve the traffic perception accuracy of connected and autonomous vehicles. The final imputation with I-GAIN further improves the accuracy of each algorithm to varying degrees, as shown in Tables 3, 4, 6, and 7.

C. IMPACT OF CAVS' MARKET PENETRATION RATE

In this section, we analyzed the influence of the CAV market penetration rate on perception accuracy. The market penetration rate in the experiment varied between 0.1 and 0.8, with 0.05 per change. The experimental results shown in Fig. 12 indicate that the perception accuracy increased as the CAV penetration rate increased. To ensure the reliability of the experimental results, we fixed the CAV permeability separately and ran the Experiment 10 times, taking the average to obtain the experimental results at that permeability. The process was repeated for each market penetration rate to

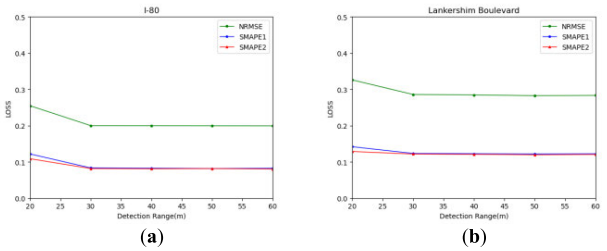

FIGURE 13. Traffic sensing accuracy under different CAV detection ranges. (a) The accuracy on I-80. (b) The accuracy on Lankershim Boulevard.

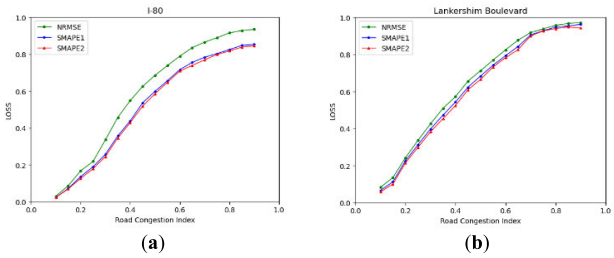

FIGURE 14. Traffic sensing accuracy under different road congestion indices. (a) Accuracy on I-80. (b) Accuracy on Lankershim Boulevard.

obtain a comprehensive analysis of the influence of the CAV market penetration rate on perception accuracy.

Fig. 12 shows that the higher the market penetration rate of CAVs, the higher the perception accuracy on both I-80 and Lankershim Boulevard. This result occurs because as more CAVs are introduced into the traffic system, there is increased communication and cooperation among the vehicles, which leads to a more accurate perception of the traffic situation. The critical market penetration rate for I-80 is approximately 0.7, while that of Lankershim Boulevard is approximately 0.75; thus, once the market penetration rate reaches these values, there is a marked improvement in perception accuracy.

However, the error of Lankershim Boulevard is larger than that of I-80 under the same market penetration rate. This result likely occurs because Lankershim Boulevard is an urban road with a denser vehicle distribution and shorter distance between vehicles. As a result, the corresponding R matrix has more rows, which increases the chance of errors in the data.

Results thus suggest that increasing the market penetration rate of CAVs can markedly improve perception accuracy, but the effect may vary depending on road characteristics such as vehicle density and road length.

D. IMPACT OF CAVS DETECTION RANGE

In this section, the influence of CAV detection range on perception accuracy is analyzed. In the experiment, the perception range of the sensor is 20-60 m, which is defined in [39]. Experimental results are shown in Fig. 13.

As shown in Fig. 13, there is a clear trend that the perception accuracy of CAVs increases as the detection range of the CAVs' onboard sensors increases on both I-80 and Lankershim Boulevard. This result is expected because increasing the detection range allows CAVs to sense more vehicles and thus obtain more accurate traffic information.

However, the perception accuracy does not continue to increase when the detection range reaches 30 m. The reason for this result is that most adjacent vehicles on I-80 and Lankershim Boulevard are usually within 30 m of each other; thus, increasing the detection range beyond this range does not markedly improve the perception accuracy.

Thus, experiments show that increasing the detection range of CAV onboard sensors can improve the perception accuracy of CAVs. However, there is a threshold for maximum improvement, and once the detection range reaches a certain level, additional increases in the detection range do not necessarily lead to improvements in perception accuracy.

E. IMPACT OF ROAD CONGESTION INDEX

In this section, the performance analysis of the CAV sensing algorithm for different road congestion indices on the same road section is reported. where the congestion index is calculated by dividing the current average real vehicle speed of the road section by the road design speed. Experimental results are shown in Fig. 14.

The experimental results shown in Fig. 14 indicate that the accuracy of the CAV traffic sensing algorithm varies at different levels of road congestion. Specifically, as the road congestion index increases, the LOSS gradually increases, and the accuracy of the algorithm decreases. These results likely occur because higher congestion levels result in more complex traffic patterns and larger areas of occlusion, which can make it more difficult for the perception algorithm to accurately detect and track individual vehicles.

Overall, the performance of CAV perception algorithms depends heavily on the specific road conditions and traffic patterns encountered, and different algorithms may be better suited for different types of roads or driving environments. Additional research is required to better understand the factors that affect the accuracy and effectiveness of CAV perception algorithms and to develop more robust and reliable methods for detecting and tracking vehicles in real-world driving scenarios.

VI. CONCLUSION

With the rapid development of CAV technology, many traffic problems can be solved using CAVs. CAVs as mobile sensors have a lot of potential to reduce or even eliminate the need for fixed-location sensors in existing transportation systems, thereby reducing costs for public agencies. However, when the market penetration rate of CAVs is low, CAVs may not be able to perceive information about all vehicles on the road.

In this study, we developed a traffic-sensing model that improves CAV perception. The model estimates the vehicle information in the perceptual blind spots using the proposed I-GAIN. To facilitate data imputation, we first model traffic states with matrices. In this process, we transform roads into road matrices based on information such as the length of vehicles, the minimum spacing between vehicles, and the number

of lanes. Then, imputation is performed, which is divided into initial and final imputations, and optimizes the GAIN algorithm into the I-GAIN algorithm to improve accuracy. Compared with other algorithms, the accuracy of the proposed I-GAIN is higher. NGSIM data were used to verify the accuracy and robustness of the proposed algorithm. Although experiments were conducted only on I-80 and Lankershim Boulevard, the proposed algorithm can be extended to any road. In addition, the effects of market penetration rate of CAVs and the detection range of sensors are also investigated.

This study and its methods have certain limitations. We only used numerical experiments to verify algorithm accuracy. Although NGSIM data are collected from pure HDV traffic flow, the characteristics of traffic flow mixed with HDVs and CAVs may have some differences. We also only used speed to verify the accuracy of the proposed method. Distance headway, time headway, and density of the mixed traffic flow also play important roles in traffic management and control for mixed traffic flow. The proposed method also did not consider weather conditions; in adverse weather conditions, the proposed method should be verified and evaluated in more detail.

REFERENCES

[1] B. Gong, R. Wei, D. Wu, and C. Lin, "Fleet management for HDVs and CAVs on highway in dense fog environment," *J. Adv. Transp.*, vol. 2020, Aug. 2020, Art. no. 8842730, doi: 10.1155/2020/8842730.

[2] S. Bahrami and M. J. Roorda, "Optimal traffic management policies for mixed human and automated traffic flows," *Transp. Res. A, Policy Pract.*, vol. 135, pp. 130–143, May 2020, doi: 10.1016/j.tra.2020.03.007.

[3] J. G. Wardrop and G. Charlesworth, "A method of estimating speed and flow of traffic from a moving vehicle," *Proc. Inst. Civil Eng.*, vol. 3, no. 1, pp. 158–171, Jan. 1954, doi: 10.1680/ipeds.1954.11628.

[4] B. Suh, Y. Shao, and Z. Sun, "Vehicle speed prediction for connected and autonomous vehicles using communication and perception," in *Proc. Amer. Control Conf. (ACC)*, Jul. 2020, pp. 448–453.

[5] S. Gupta and M. Canova, "Eco-driving of connected and autonomous vehicles with sequence-to-sequence prediction of target vehicle velocity," in *Proc. 6th IFAC Conf. Engine Powertrain Control, Simulation Modeling (E-COSM)*, Aug. 2021, pp. 430–436, doi: 10.1016/j.ifacol.2021.10.200.

[6] H. Li and W. Li, "Estimating the average road travel time based on soft set under connected and autonomous vehicles," in *Proc. 5th Int. Conf. Inf. Sci., Comput. Technol. Transp. (ISCTT)*, Nov. 2020, pp. 566–570, doi: 10.1109/ISCTT51595.2020.00108.

[7] Y. Shao and Z. Sun, "Eco-approach with traffic prediction and experimental validation for connected and autonomous vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1562–1572, Mar. 2020, doi: 10.1109/TITS.2020.2972198.

[8] W. Ma and S. Qian, "High-resolution traffic sensing with probe autonomous vehicles: A data-driven approach," *Sensors*, vol. 21, no. 2, p. 464, Jan. 2021, doi: 10.3390/s21020464.

[9] J. Yoon, J. Jordon, and M. Schaar, "GAIN: Missing data imputation using generative adversarial nets," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5689–5698.

[10] V. Alexiadis, J. Colyar, J. Halkias, R. Hranac, and G. McHale, "The next generation simulation program," *ITE J. Inst. Transp. Eng.*, vol. 74, no. 8, p. 22, 2004.

[11] Y. Wei, Q. Tian, J. Guo, W. Huang, and J. Cao, "Multi-vehicle detection algorithm through combining Harr and HOG features," *Math. Comput. Simul.*, vol. 155, pp. 130–145, Jan. 2018, doi: 10.1016/j.matcom.2017.12.011.

[12] M.-S. Wang and Z.-R. Zhang, "FPGA implementation of HOG based multi-scale pedestrian detection," in *Proc. IEEE Int. Conf. Appl. Syst. Invention (ICASI)*, Apr. 2018, pp. 1099–1102, doi: 10.1109/ICASI.2018.8394472.

[13] Q. Xu, X. Zhang, R. Cheng, Y. Song, and N. Wang, "Occlusion problem-oriented adversarial faster-RCNN scheme," *IEEE Access*, vol. 7, pp. 170362–170373, 2019, doi: 10.1109/ACCESS.2019.2955685.

[14] R. H. Rasshofer and K. Gresser, "Automotive radar and LiDAR systems for next generation driver assistance functions," *Adv. Radio Sci.*, vol. 3, pp. 205–209, May 2005, doi: 10.5194/ars-3-205-2005.

[15] H. Liu, C. Lin, B. Gong, and D. Wu, "Extending the detection range for low-channel roadside LiDAR by static background construction," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5702412, doi: 10.1109/TGRS.2022.3155634.

[16] C. Lin, G. Sun, L. Tan, B. Gong, and D. Wu, "Mobile LiDAR deployment optimization: Towards application for pavement marking stained and Worn detection," *IEEE Sensors J.*, vol. 22, no. 4, pp. 3270–3280, Feb. 2022, doi: 10.1109/JSEN.2022.3140312.

[17] C. Lin, Y. Guo, W. Li, H. Liu, and D. Wu, "An automatic lane marking detection method with low-density roadside LiDAR data," *IEEE Sensors J.*, vol. 21, no. 8, pp. 10029–10038, Apr. 2021, doi: 10.1109/JSEN.2021.3057999.

[18] J. Zhao, H. Xu, H. Liu, J. Wu, Y. Zheng, and D. Wu, "Detection and tracking of pedestrians and vehicles using roadside LiDAR sensors," *Transp. Res. C, Emerg. Technol.*, vol. 100, pp. 68–87, Mar. 2019, doi: 10.1016/j.trc.2019.01.007.

[19] Y. Zhang, N. Bhattarai, J. Zhao, H. Liu, and H. Xu, "An unsupervised clustering method for processing roadside LiDAR data with improved computational efficiency," *IEEE Sensors J.*, vol. 22, no. 11, pp. 10684–10691, Jun. 2022, doi: 10.1109/JSEN.2022.3166957.

[20] J. Zheng and H. X. Liu, "Estimating traffic volumes for signalized intersections using connected vehicle data," *Transp. Res. C, Emerg. Technol.*, vol. 79, pp. 347–362, Jun. 2017, doi: 10.1016/j.trc.2017.03.007.

[21] T. Li, X. Han, and J. Ma, "Cooperative perception for estimating and predicting microscopic traffic states to manage connected and automated traffic," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 13694–13707, Aug. 2022, doi: 10.1109/TITS.2021.3126621.

[22] S. Wei, D. Yu, C. L. Guo, L. Dan, and W. W. Shu, "Survey of connected automated vehicle perception mode: From autonomy to interaction," *IET Intell. Transp. Syst.*, vol. 13, no. 3, pp. 495–505, 2018, doi: 10.1049/iet-its.2018.5239.

[23] C. M. Day and D. M. Bullock, "Detector-free signal offset optimization with limited connected vehicle market penetration: Proof-of-concept study," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2558, no. 1, pp. 54–65, Jan. 2016, doi: 10.3141/2558-06.

[24] D. Li and P. Wagner, "A novel approach for mixed manual/connected automated freeway traffic management," *Sensors*, vol. 20, no. 6, p. 1757, Mar. 2020, doi: 10.3390/s20061757.

[25] J. W. Graham, "Missing data analysis: Making it work in the real world," *Annu. Rev. Psychol.*, vol. 60, pp. 549–576, Jan. 2009, doi: 10.1146/annurev.psych.58.110405.085530.

[26] J. Z. S. Zhang, X. Zhu, Y. Qin, and C. Zhang, "Missing value imputation based on data clustering," in *Transactions on Computational Science I* (Lecture Notes in Computer Science), 2008, vol. 4750, no. 1. Berlin, Germany: Springer, pp. 128–138, doi: 10.1007/978-3-540-79299-4_7.

[27] D. Ball, *Statistical Analysis With Missing Data*. Hoboken, NJ, USA: Wiley, 2003.

[28] Y. Qin, S. Zhang, X. Zhu, J. Zhang, and C. Zhang, "Semi-parametric optimization for missing data imputation," *Appl. Intell.*, vol. 27, no. 1, pp. 79–88, 2007, doi: 10.1007/s10489-006-0032-0.

[29] D. B. Rubin, "Multiple imputation after 18+ years," *J. Amer. Stat. Assoc.*, vol. 91, no. 434, pp. 473–489, Jun. 1996, doi: 10.1080/01621459.1996.10476908.

[30] G. E. Batista and M. C. Monard, "A study of K-nearest neighbour as an imputation method," *His*, vol. 87, pp. 251–260, Dec. 2002.

[31] D. J. Stekhoven and P. Buhlmann, "MissForest-non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, Jan. 2012, doi: 10.1093/bioinformatics/btr597.

[32] L. Gondara and K. Wang, "MIDA: Multiple imputation using denoising autoencoders," 2017, *arXiv:1705.02737*.

[33] J. T. McCoy, S. Kroon, and L. Auret, "Variational autoencoders for missing data imputation with application to a simulated milling circuit," *IFAC-PapersOnLine*, vol. 51, no. 21, pp. 141–146, 2018, doi: 10.1016/j.ifacol.2018.09.406.

[34] A. Nazábal, P. M. Olmos, Z. Ghahramani, and I. Valera, "Handling incomplete heterogeneous data using VAEs," *Pattern Recognit.*, vol. 107, Nov. 2020, Art. no. 107501, doi: 10.1016/j.patcog.2020.107501.

[35] M. Ramezani, J. A. Machado, A. Skabardonis, and N. Geroliminis, "Capacity and delay analysis of arterials with mixed autonomous and human-driven vehicles," in *Proc. 5th IEEE Int. Conf. Models Technol. Intell. Transp. Syst. (MT-ITS)*, Jun. 2017, pp. 280–284, doi: 10.1109/MTITS.2017.8005680.

[36] J. Van Brummelen, M. O'Brien, D. Gruyer, and H. Najjaran, "Autonomous vehicle perception: The technology of today and tomorrow," *Transp. Res. C, Emerg. Technol.*, vol. 89, pp. 384–406, Apr. 2018, doi: 10.1016/j.trc.2018.02.012.

[37] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001, doi: 10.1093/bioinformatics/17.6.520.

[38] S. H. Huang, R. Kothamasu, and N. Rapur, "Iterative imputation algorithms for process modeling with incomplete data," *Intell. Data Anal.*, vol. 11, no. 2, pp. 189–202, Apr. 2007, doi: 10.3233/IDA-2007-11206.

[39] R. Sun, "Matrix completion via nonconvex factorization: Algorithms and theory," Ph.D. dissertation, Dept. Elect. Comput. Eng., Univ. Minnesota, Minneapolis, MN, USA, 2015.

BOWEN GONG received the Ph.D. degree in traffic information and control engineering from Jilin University, China, in 2010. She is currently an Associate Professor with the Department of Traffic Information and Control Engineering, Jilin University. Her research interests include traffic perception, intelligent traffic navigation, traffic behavior analysis, traffic safety evaluation, and prevention.

ZHIPENG XU received the bachelor's degree in traffic engineering from the Ningbo University of Technology, China, in 2020. He is currently pursuing the master's degree in traffic and transportation engineering with Jilin University, China. His research interests include traffic perception and mixed traffic flow control.

CIYUN LIN received the Ph.D. degree in traffic information and control engineering from Jilin University, China, in 2010. He is currently a Professor with the Department of Traffic Information and Control Engineering, Jilin University. His research interests include traffic perception, intelligent transportation systems, traffic operation and control, and traffic data analytics.

DAYONG WU received the Ph.D. degree in transportation engineering from Texas Tech University, Lubbock, TX, USA, in 2008. He is currently an Assistant Research Scientist with the Texas A&M Transportation Institute, Texas A&M University, USA. His research interests include intelligent transportation systems, GIS in transportation, and traffic safety.