PeerJ Computer Science

Traffic data imputation *via* knowledge graph-enhanced generative adversarial network

Yinghui Liu[1], Guojiang Shen[1], Nali Liu[1], Xiao Han[2], Zhenhui Xu[3], Junjie Zhou[3] and Xiangjie Kong[1]

[1] College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China
[2] School of Data Science, City University of Hong Kong, Hong Kong, China
[3] Zhejiang Supcon Information Co., Ltd., Hangzhou, China

ABSTRACT

Traffic data imputation is crucial for the reliability and efficiency of intelligent transportation systems (ITSs), forming the foundation for downstream tasks like traffic prediction and management. However, existing deep learning-based imputation methods struggle with two significant challenges: poor performance under high missing data rates and the limited incorporation of external traffic-related factors. To address these challenges, we propose a novel knowledge graph-enhanced generative adversarial network (KG-GAN) for traffic data imputation. Our approach uniquely integrates external knowledge with traffic spatiotemporal dependencies to improve data imputation quality. Specifically, we construct a fine-grained knowledge graph (KG) that differentiates attributes and relationships of external factors such as points of interest (POI) and weather conditions, facilitating more robust knowledge representation learning. We then introduce a knowledge-aware embedding cell (EM-cell) that merges traffic data with these learned external representations, providing richer inputs for the spatiotemporal GAN. Extensive experiments on a large-scale real-world traffic dataset demonstrate that KG-GAN significantly outperforms state-of-the-art methods under various missing data scenarios. Additionally, ablation studies confirm the superior performance gained from incorporating external knowledge, underscoring the importance of this approach in addressing complex missing data patterns.

Subjects Artificial Intelligence, Data Mining and Machine Learning
Keywords Traffic data imputation, Generative adversarial networks, Knowledge graph

Submitted 3 June 2024
Accepted 22 September 2024
Published 14 October 2024

Corresponding author
Xiangjie Kong, xjkong@ieee.org

Academic editor
Valentina Emilia Balas

Additional Information and Declarations can be found on page 19

DOI **10.7717/peerj-cs.2408**

INTRODUCTION

Traffic detection data collected in intelligent transport systems (ITSs) often suffer from missing data due to various technical and management issues, including software failures, power outages, transmission errors, or storage failures (*Tan et al., 2014*), as shown in Fig. 1. For example, the Caltrans performance measurement system (PEMS) can be used to collect traffic data, calculate highway usage and congestion delays, predict travel time, evaluate ramp metering methods, and validate traffic theories. However, the data samples received by PEMS are often incomplete. According to the statistics, the ITS in Beijing, China was still under development in 2008, the daily traffic flow data had a general loss rate

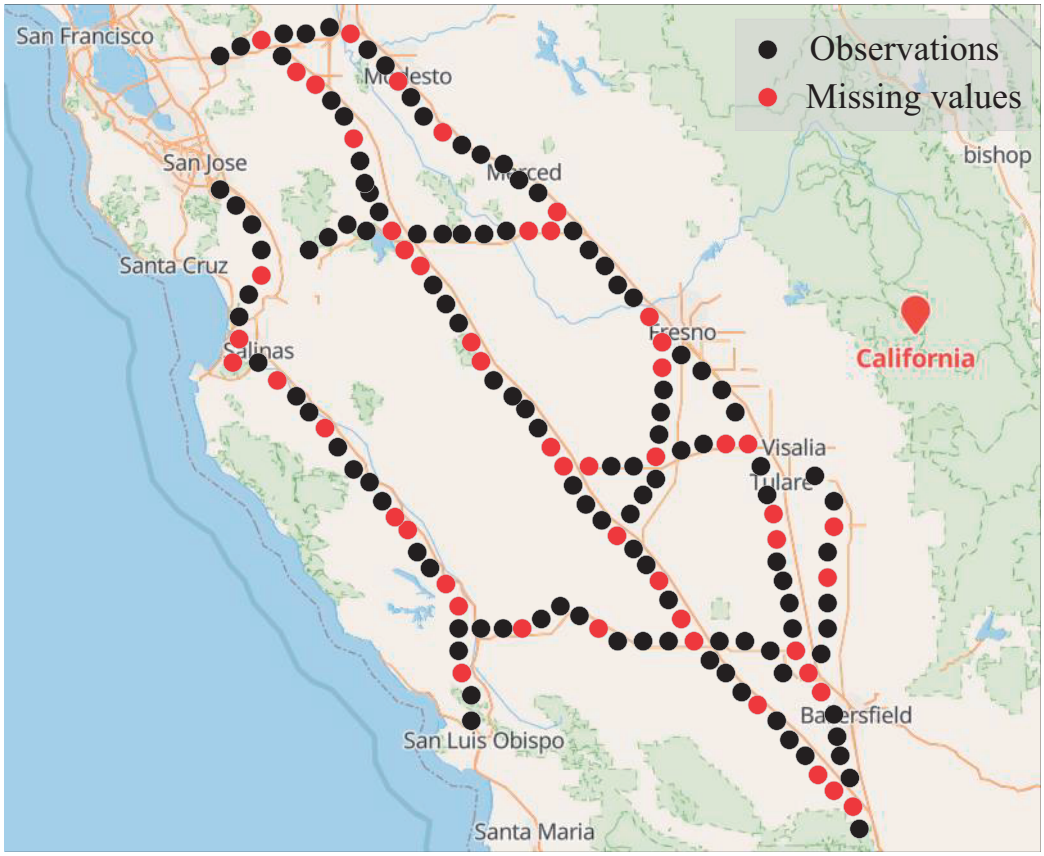

Figure 1 **Road network sensors with missing data.** Map data © 2024 Google.
Full-size DOI: 10.7717/peerj-cs.2408/fig-1

of around 10% (4% due to detector failure, 6% due to other reasons), and some loop detectors even generated a missing rate as high as 20–25% (*Qu et al., 2009*). Some extreme missing scenarios were reported in Alberta, Canada. *Jianrui, Xingyi & Yi (2010)* pointed out that nearly 50% of traffic data was missing in 7 years. Missing data seriously affects the accuracy and reliability of traffic forecasting (*Olayode et al., 2024*), management, and control systems (*Kong et al., 2024*; *Pamuła, 2018*). To address the missing data problem, data imputation is crucial to reconstruct the dataset by filling in the missing values with robust estimates.

Most previous imputation methods for traffic data fall into three categories: 1) Traditional statistical methods, such as support vector regression (SVR) (*Wu, Ho & Lee, 2004*), autoregressive integrated moving average model (ARIMA) (*Cetin & Comert, 2006*), mean imputation, median imputation (*Kaiser, 2014*) and other algorithms (*Bania & Halder, 2020*; *Caillault, Lefebvre & Bigand, 2020*). These methods rely on smoothness and periodicity to interpolate missing values. However, there is uncertain variation in real life, which leads to the unsatisfactory results of these methods; 2) Tensor decomposition methods (new machine learning-based methods) (*Chen, He & Sun, 2019*; *Zhang et al., 2021*; *Chen et al., 2023*, *2024*). This category of methods estimates the missing values in the traffic flow by obtaining a suitable low-rank approximation of the incomplete matrix. 3)

Deep learning methods (*Kong et al., 2023*; *Ni & Cao, 2022*; *Shen et al., 2023*; *Tan et al., 2020*; *Tian et al., 2018*). They interpolate data by learning the temporal and spatial correlation of data or the distribution of data.

However, deep learning methods achieve inferior performance compared to the tensor decomposition methods under high data missing rate. In addition, these methods are only limited to imputing data with traffic data itself. In addition to being influenced by the quality of road detectors and the distribution of spatiotemporal features, traffic information may also be affected by various external factors, which is the efficient promotion of knowledge-driven data imputation. For example, weather conditions, the existence of traffic stations, emergencies, holidays, and the distribution of nearby Points of Interest (POIs) (*Lana et al., 2018*; *Liu et al., 2024*; *Xu et al., 2022*). These external factors may affect urban traffic data directly or indirectly. For example, the traffic volume under different weather conditions may have different states as the weather changes over time. What's more, traffic data is not only influenced by a single factor but also by various factors. For example, under the same heavy rain conditions, the traffic volume around schools is more affected than on less popular roads nearby. Integrating the semantic correlation of multi-source data information is the key to improving the ability to impute traffic data. Fortunately, knowledge graphs (KGs) (*Li et al., 2023*; *Peng et al., 2023*) that contain rich semantics about entities and relations provide a way to integrate different external factors and represent them in a unified manner.

In light of the above limitations and challenges, we propose a knowledge graph-enhanced generative adversarial network (KG-GAN) for spatiotemporal traffic data imputation. This approach is designed to address the complex nature of traffic data, which is influenced not only by temporal and spatial dependencies but also by various external factors such as points of interest (POI), weather conditions, and other contextual information. To tackle the challenge of effectively incorporating external factors, we use a traffic-specific KG construction approach that distinguishes the attributes and relationships of external entities. This allows the model to capture fine-grained semantic correlations that are often overlooked in traditional methods. The constructed KG serves as a rich source of prior knowledge, enhancing the representation learning process. After learning these knowledge representations, we introduce a knowledge-aware embedding cell (EM-cell). This component is specifically designed to seamlessly integrate the learned KG representations with traffic data, enriching the traffic embeddings with semantic information. These enriched embeddings are then fed into a spatiotemporal generative adversarial network, which is capable of generating high-quality imputed data by effectively modeling both the spatiotemporal dependencies and the complex external correlations. Compared to tensor decomposition methods that primarily focus on reducing the dimensionality of high-dimensional data, our KG-GAN approach benefits from the integration of prior knowledge in the form of KGs. This not only compensates for the limitations of purely data-driven deep learning methods but also enhances the model's adaptability to complex and diverse missing data patterns, ultimately leading to more robust and accurate imputation results.

The following are our summarized contributions:

- To improve the performance of the deep learning imputation model while considering the complex influence of external factors on traffic interpolation, we propose a KG-enhanced approach (namely KG-GAN), in which traffic spatio-temporal characteristics and external knowledge graph are jointly learned.
- We design a knowledge-aware embedding cell (EM-cell) to enrich model inputs by integrating traffic data with external knowledge, in which we construct implicit knowledge representations of external factors with a fine-grained KG construction approach that distinguishes the attributes and relations of external entities.
- To demonstrate the effectiveness of KG-GAN, we conduct extensive experiments on a large-scale real-world traffic dataset showing that our method significantly outperforms existing imputation models and enhances performance across various missing data patterns. Further ablation experiments are performed to highlight the superiority of incorporating external knowledge learning.

The rest of this article is organized as follows. "Related Work" provides a systematic review of related works. "Methodology" describes the architecture and details of the proposed KG-GAN model. "Experiments" discusses the results of the experiment. Finally, the article is summarized in "Conclusion".

RELATED WORK

Data imputation methods based on deep learning

Deep learning has been successfully applied in the field of data imputation. *Che et al. (2018)* proposed a deep model (GRU-D) based on learning gated recurrent units (GRU) (*Cho et al., 2014*). GRU-D employs two distinct representations of missing patterns, namely masking and time interval. Where the masking representation simulates the location of missing data, while the time interval representation represents the time range from the last observed value. GRU-D effectively integrates them into the deep model architecture. As a result, it can capture long-term temporal dependencies in time series. Furthermore, *Cao et al. (2018)* proposed a recurrent neural network (RNN) based model for the imputation of missing data in time series (BRITS). BRITS is used to interpolate missing values in time series data and can learn missing values directly in a bidirectional RNN without any specific preprocessing. It treats the interpolated values as variables of the RNN and can be efficiently updated during backpropagation. Both GRU-D and BRITS models are based on RNN, but they only consider the temporal correlation of data and do not consider the effect of spatial information on the imputed road network data. Generative adversarial networks (GANs) have been widely used in image processing (*Xu et al., 2018*; *Yi, Walia & Babyn, 2019*), and in recent years, they have been found to have good performance in data imputation. *Yoon, Jordon & Schaar (2018)* proposed a GAN-based GAIN model where they used generator and discriminator adversarial learning to model the distribution of the original data and then achieve the effect of imputing the missing data. *Wang et al. (2021)* proposed a PC-GAIN model which added a GAIN-based

pre-training process. However, these GAN-based models do not take into account the spatial and temporal correlation of the data, resulting in unsatisfactory imputation of the traffic data. *Ye, Zhang & Yu (2021)* proposed a graph attention network model (GACN) for traffic missing data imputation, which follows an encoder-decoder structure and introduces a graph attention mechanism to learn the traffic graph. It allows higher-quality traffic data to be estimated by extracting typical spatiotemporal features. Different from previous works, we use a multi-perspective spatiotemporal generative adversarial network to analyze and extract traffic features from three perspectives: temporal, spatial, and spatiotemporal (*Li et al., 2018*; *Shen et al., 2022*).

Knowledge representation of traffic data

The generation of multi-source data is a natural consequence of a complex hybrid urban transport system. The relationships in multi-source data are mainly presented as networks, and mining the structural and relational information contained in the networks through representation vectors becomes the main method to capture the network information. In general, networks can be classified into homogeneous and heterogeneous networks based on the type of nodes. Most realistic traffic states are heterogeneous network structures, but the traditional HEBE (*Gui et al., 2017*) embedding framework for handling heterogeneous networks is only adapted to specific network architectures due to the limitation of meta-path accuracy. In recent years the application of knowledge graphs has gradually entered the public domain, and they are used in the traffic field for their excellent ability to handle graph structures and information (*Muppalla et al., 2017*; *Xu et al., 2016*). Typical knowledge graph representation learning methods include TransE (*Bordes et al., 2013*), TransH (*Wang et al., 2014*), and TransR (*Lin et al., 2015*). Compared to TransE and TransH, TransR constructs a projection matrix to model entities and relationships in both entity space and relationship space, and performs translation in the relationship space, breaking the limitation of the same space. Therefore, we choose TransR to model our constructed entities, relationships, and attributes.

METHODOLOGY

Overall framework

Our proposed KG-GAN model perceives the semantic information of the external knowledge graph and the spatiotemporal relevance of the traffic features through adversarial learning, thus effectively improving the accuracy of traffic data imputation. The model architecture diagram is shown in Fig. 2. The architecture takes in road network data, traffic speed data, and knowledge graphs to facilitate the learning of semantic knowledge and spatiotemporal dependencies, ultimately generating imputed data. In particular, following *Lin, Liu & Sun (2016)*, we first divide the knowledge triad into a relation triad and an attribute triad to realize the refinement of the knowledge attribute and relationship. Then, we use the knowledge representation model with entities, attributes and relations (KR-EAR) to train the triad to generate the relational representation matrix to characterize the implicit knowledge. In addition, we propose a knowledge-aware embedding cell (EM-Cell) to fuse the implicit representation of the knowledge graph with traffic features

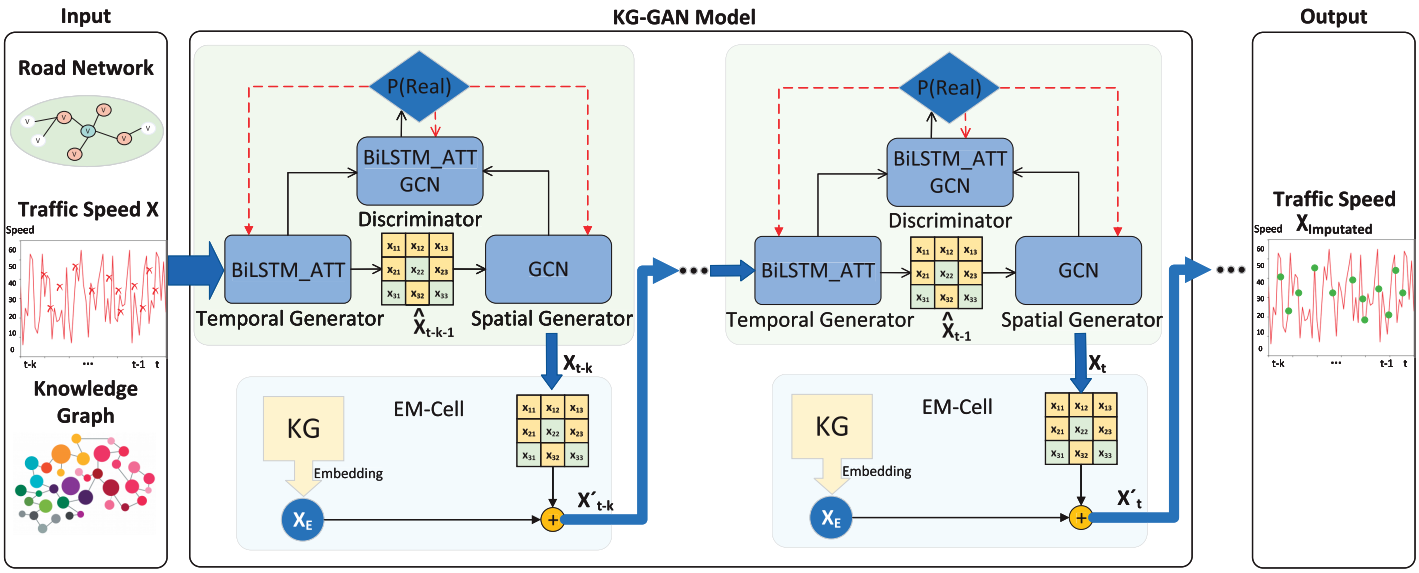

Figure 2 **The KG-GAN model framework.** Full-size DOI: 10.7717/peerj-cs.2408/fig-2

for nonlinear self-learning, and input the learned-well traffic embedding with rich semantic information into our previously proposed multi-perspective spatiotemporal generative adversarial network (MST-GAN) (*Shen et al., 2022*) to guide the convergence and optimization of the model. Ultimately, it enables the model to interpolate data with complex correlations between traffic spatiotemporal features and external factors.

Given the incomplete traffic observed data X, the traffic imputation problem can be considered to learn an imputation function *Func*, which can calculate an appropriate value for each missing component in X based on the traffic network structure matrix A and the knowledge graph (KG) as follows:

$$X_{imputed} = Func(A, X, KG). \tag{1}$$

Combined with the adversarial training of the MST-GAN model, the final min-max objective for the overall model optimization is:

$$\min_{G_T, G_S} \max_{D} E_{\bar{X},M}[M \cdot \log D(\bar{X}) + (1 - M) \cdot \log(1 - D(\bar{X})], \tag{2}$$

where G_T characterizes the temporal generator, G_S characterizes the spatial generator, D characterizes the discriminator, $\bar{X}$ denotes the road network data simulated by the temporal and spatial generator, M characterizes the masking matrix, and $\log(\cdot)$ characterizes the logarithmic calculation of the elements.

The design and learning of knowledge graph

KG is a semantic network-based knowledge base that uses a directed graph structure to organize data such as entities, relationships, and attributes. The advantages of KG, such as their ability to integrate diverse information sources and preserve both semantic and

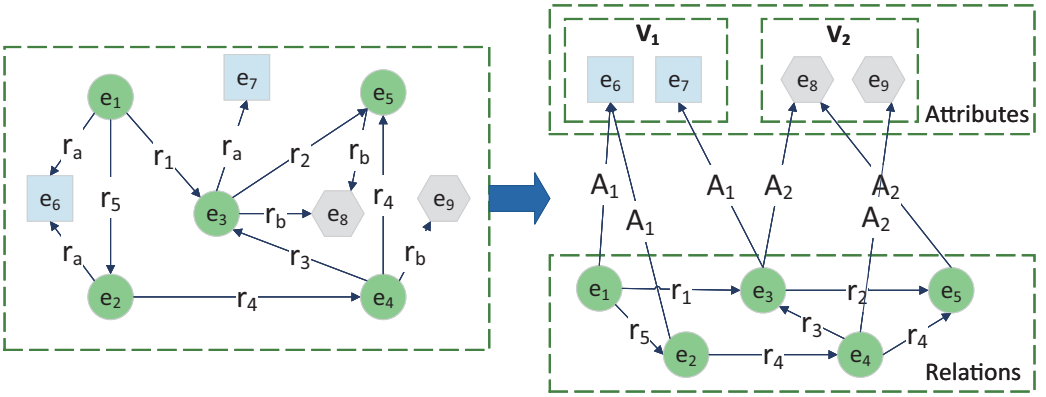

Figure 3 **The KR-EAR (right) and traditional KR method (left).**
Full-size DOI: 10.7717/peerj-cs.2408/fig-3

structural relationships, are particularly beneficial for traffic data imputation. In handling incomplete traffic data, KG can effectively model the complex, multi-relational dependencies inherent in traffic networks. By representing heterogeneous nodes and multi-relationship information, KG can construct hierarchical and semantic relationships between various traffic-related entities (*e.g.*, road segments, sensors, events) (*Ning et al., 2024*). This hierarchical and semantic structuring allows for more accurate imputation by leveraging the rich contextual and relational information within the graph, which is critical in capturing the dynamic and interdependent nature of traffic systems.

Distributed knowledge representation (KR) encodes entities and relations in a low-dimensional semantic space, significantly improving the performance of relation extraction and knowledge inference. In many KGs, some relations represent attributes of entities (properties), while others represent relationships between entities (relations). Traditional KR methods treat all relations equally and usually have poor accuracy in modeling one-to-many and many-to-one relations (consisting mainly of attributes). In principle, a knowledge graph representation that distinguishes between attribute and relationship information is more suitable for capturing semantic information and relevance in this context. Therefore, we use the knowledge graph representation method knowledge representation learning with entities, attributes and relations (KR-EAR) (*Lin, Liu & Sun, 2016*) based on entity-attribute and entity-relationship to capture knowledge structure and semantic information between road parts and external factors. The KR-EAR and the traditional KR method are shown specifically in Fig. 3, where A_1 and A_2 are the two attributes. The value set of attribute $A_1(V_1)$ contains e_6 and e_7 which are squares (also colored in blue), while $A_2(V_2)$ contains e_8 and e_9 which are hexagonal (also colored in grey). In the traditional KR representation method (left), attributes A_1 and A_2 are treated as relations r_a and r_b. In contrast, KR-EAR encodes the relational triples using the traditional KR representation method and treats attribute prediction as a classification problem.

In this article, roads, attributes, and the relationships between them are represented as a triad of KG = {*R*,*ATT*, *Relations*}. Specifically, the triads are divided into three categories:

1) **Road adjacency triple** R (head entity, relationship, tail entity)

$$R = \{(v_i, adj, v_j)\}, i, j \in \{1, 2, \ldots, n\}, \tag{3}$$

where R is a relational triplet representing the adjacency relationship adj between segments v_i and v_j, and n is the number of segments.

2) **Attribute triple** ATT (entity, attribute, attribute value)

$$ATT = \{(v_i, a_l, av_{a_l})\}, l \in \{1, 2, \ldots, L\}, \tag{4}$$

where a_l is the l-th class of attributes, av_{a_l} is the corresponding attribute value (*e.g.*, weather overcast), and L is the number of attribute classes.

3) **Attribute co-occurrence triple** *Relations* (attribute 1, attribute 2, co-occurrence probability)

$$Relations = \{(a_{l_1}, a_{l_2}, p)\}, l_1, l_2 \in \{1, 2, \ldots, L\}, \tag{5}$$

where a_{l_1} and a_{l_2} denote two different attributes of an entity, p is their co-occurrence probability and the attribute co-occurrence probability describes the probability that two attributes exist in the same section.

Given a KG, the objective of KR-EAR is to learn the representations X_E of entities, relations, and attributes. The objective function is defined as maximizing the joint conditional probability of the relationship triple and the attribute triple, which is formalized as:

$$\begin{aligned} P(R, ATT|X_E) &= P(R|X_E)P(ATT|X_E), \\ &= \prod_{(v_i, adj, v_j) \in R} P((v_i, adj, v_j)|X_E) \prod_{(v_i, a_l, av_{a_l}) \in ATT} P((v_i, a_l, av_{a_l})|X_E), \end{aligned} \tag{6}$$

where $P((v_i, adj, v_j)|X_E)$ denotes the conditional probability of the relation triple (v_i, adj, v_j) and $P((v_i, a_l, av_{a_l})|X_E)$ is the conditional probability of the attribute triple (v_i, a_l, av_{a_l}). $P((v_i, adj, v_j)|X_E)$ is generated by an energy function e following TransR (*Lin et al., 2015*):

$$P((v_i, adj, v_j)|X_E) = \frac{exp(e(v_i, adj, v_j))}{\sum_{\hat{v}_i \in V} exp(e(\hat{v}_i, adj, v_j))}, \; e(h, r, t) = -||hM_r + r - tM_r||_{L_1/L_2} + b_r, \tag{7}$$

where M_r denote the projection matrix which may projects entities from entity space to relation space. b_r is a bias constant and V is a set of road section entities. $P((v_i, a_l, av_{a_l})|X_E)$ is captured by a scoring function (*Lin, Liu & Sun, 2016*):

$$P((v_i, a_l, av_{a_l})|X_E) = \frac{exp(s(v_i, a_l, av_{a_l}))}{\sum_{\widehat{av}_{a_l} \in AV_{a_l}} exp(s(v_i, a_l, \widehat{av}_{a_l}))}, \tag{8}$$

where $s()$ is the scoring function for each attribute value of a given entity and AV_{a_l} is the attribute value set. In this way, KR-EAR generates representations of relations and attributes while strengthening the correlations between attributes.

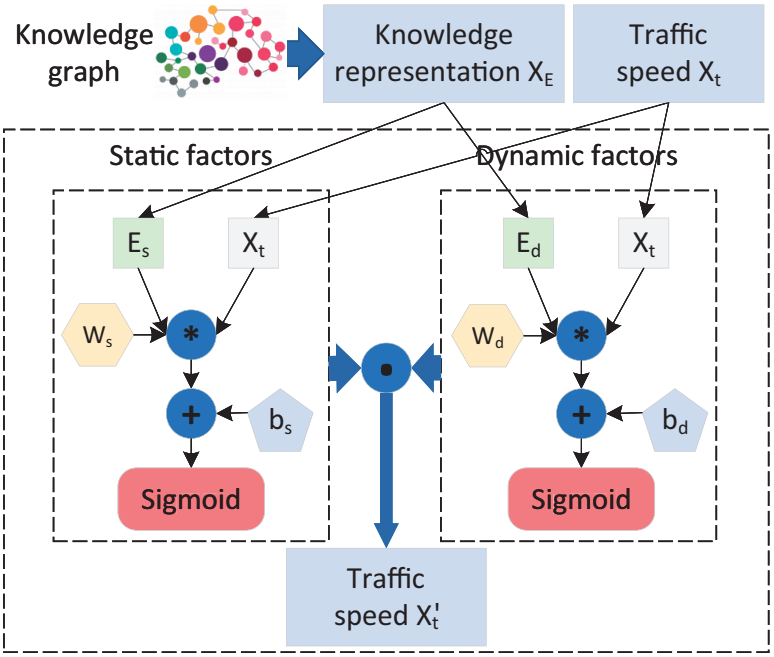

Figure 4 **Structure of EM-cell.** Full-size DOI: 10.7717/peerj-cs.2408/fig-4

The fusion of knowledge representations and GAN

In order to better model the spatiotemporal dependencies of traffic data and perceive the influences of external factors from multiple perspectives, as well as the correlations between factors, this study proposes a knowledge-aware embedding method, namely EM-Cell. The design of EM-Cell is based on a deep analysis of traffic data and the mining of multi-source knowledge relationships, which can effectively fuse the complex knowledge representations of spatiotemporal changes of traffic data and external factors. The details of EM-Cell are shown in Fig. 4, where the input consists of two parts: the knowledge representation matrix X_E constructed by KR-EAR and the road segment feature X_t observed at time t. Due to the diversity of external factors, this article divides them into two categories: static factors and dynamic factors. Specifically, E_s and E_d in Fig. 4 represent the embeddings of road segments with respect to static external factors (such as shopping mall information, hospital information) and dynamic external factors (such as weather changes), respectively, processed by KR-EAR. Therefore, the fusion operation formula between the traffic feature matrix and the knowledge representation matrix designed in this study is as follows:

$$X_t' = Concat[\sigma(f_s(E_s, X_t)), \sigma(f_d(E_d, X_t))], \ \ where\, f(x, y) = xyW + b, \ f = \{f_s, f_d\}. \tag{9}$$

Both $W = \{W_s, W_d\}$ and $b = \{b_s, b_d\}$ denote the learnable parameters. σ is the *sigmoid* function.

To model the spatiotemporal dependence of traffic data based on knowledge representation, we use the updated road segment features X_t' and the adjacency matrix A as the input to the spatiotemporal generative adversarial network. We use the MST-GAN

model for data imputation because it considers advanced multi-view spatiotemporal fusion through chain generator adversarial learning. To achieve multi-view feature fusion, MST-GAN uses an adversarial between a chain generator and a discriminator to achieve a high-level fusion of temporal and spatial information. The generator learns different enhanced features flexibly at different stages using independent parameters. In summary, MST-GAN captures the temporal and spatial correlation of traffic data through a bidirectional recurrent network and a graph convolutional network. In addition, it introduces an attention layer to compute dynamic weights between different time points to focus on key temporal features.

Specifically, the temporal generator G_T uses a bi-directional long and short-term memory network based on the attention mechanism ($BiLSTM\_ATT$) as the kernel, and the final output value of the temporal generator $\hat{X}$ is denoted as:

$$X_{BiLSTM\_ATT} = Attention(BiLSTM(X'_t)), \tag{10}$$

$$\hat{X} = X'_t \odot M + X_{BiLSTM\_ATT} \odot (1 - M), \tag{11}$$

where $\odot$ means multiplying by elements and M represents the mask matrix. The spatial generator G_S kernel consists of a graph convolutional network (GCN), containing two convolutional layers and a fully connected layer. As the depth of the network increases, it brings many problems such as gradient dissipation. Therefore, the model uses skip connect to improve the gradient dissipation problem during backpropagation. The output of the spatial generator is expressed as:

$$\bar{X} = X'_t \odot M + G_S(\hat{X}) \odot (1 - M). \tag{12}$$

Both time and spatial generators use MSE as the loss function and Adam as the optimization function, the loss gradually decreases to stability during the cyclic iteration of the model. Eventually it generates interpolated data with higher quality.

The discriminator D is used to distinguish the data as true or false. The generator tries hard to make the simulated data closer to the true value, while the discriminator tries hard to identify the data as true or false. The core structure of the discriminator network consists of GCN and $BiLSTM\_ATT$. The loss function of the discriminator can be expressed as:

$$L_D = -\frac{1}{n}\sum_{i=1}^{n}(M \cdot \log D(\bar{X}) + (1 - M) \cdot \log(1 - D(\bar{X})), \tag{13}$$

where n is the sample number.

KG-GAN training for missing data imputation

The detailed training procedure of KG-GAN is presented in Algorithm 1. Firstly, refine entity attributes and entity relationships using the KR-EAR method to construct triples of correlated knowledge and derive knowledge representation matrix X_E. Then, through the EM-Cell module, fuse the spatiotemporal traffic data X_t (which may contain missing data) with the dynamic and static matrices of the knowledge representation matrix to obtain integral embedding of rich traffic information. Finally, at each training time step, the fused

Algorithm 1 **KG-GAN model training for data imputation.**

Input: Original complete traffic dataset $\boldsymbol{X_t}(m \times n)$; road adjacency triplet $\boldsymbol{R}$; attribute triplet $\boldsymbol{ATT}$; attribute co-occurrence triplet $\boldsymbol{Relations}$; loss hyperparameters $\boldsymbol{\alpha}$; masking matrix $\boldsymbol{M}$; indicator matrix $\boldsymbol{H}$; the number of epochs $\boldsymbol{N}$ and the initialized parameters: generator $\boldsymbol{\theta_{G_T}}$, generator $\boldsymbol{\theta_{G_S}}$, and discriminator $\boldsymbol{\theta_D}$

Construct a KG = {R, ATT, $Relations$} triplet.

By maximizing the $P(R, ATT|X_E)$ obtain the knowledge representation matrix X_E.

Integrating X_t and X_E : $X_t' = EM\text{-}Cell(X_t, X_E)$

for epoch=1,2,…,N

    (1) Discriminator optimization:

        Obtain discriminator loss L_D *via* Eqs. (10–13)

        Back-propagate L_D to update θ_D

    (2) Generator optimization:

        Obtain the output of $G_T\hat{X}$ *via* Eqs. (10), (11)

        $L_{R1} = \frac{1}{n}\sum_{i=1}^{n} ||(X_t' - \hat{X}) \odot M||_2^2$

        $L_{G_T} \leftarrow -D(\hat{X}, H) + \alpha L_{R1}$

        Obtain the output of $G_S\bar{X}$ *via* Eq. (12)

        $L_{R2} = \frac{1}{n}\sum_{i=1}^{n} ||(X_t' - \bar{X}) \odot M||_2^2$

        $L_{G_S} \leftarrow -D(\bar{X}, H) + \alpha L_{R2}$

        Back_propagate L_{G_T}, L_{G_S} to update θ_{G_T}, θ_{G_S}

end for

Impute the missing values:

Obtain imputed data $X_{imputed}$ *via* Eqs. (10–12)

Output: Trained parameters θ_{G_T}, θ_{G_S}, and θ_D; imputed data $X_{imputed}$

input X_t' is subjected to adversarial optimization in the MST-GAN network. During the training phase of the MST-GAN model, the discriminator is pre-trained to learn the characteristics of the generated data and observed data. Then, the discriminator and two generators are trained adversarially. In detail, the MST-GAN model uses the temporal generator to train hyperparameter θ_{G_T} from a temporal perspective. Next, we use the spatial generator to train hyperparameter θ_{G_S} from a spatial perspective. The learning of θ_{G_T} and θ_{G_S} allows us to refine the extraction of spatiotemporal features in stages.

EXPERIMENTS

Dataset

The dataset contains road network data, weather data, POI data and traffic speed data for each street in Luohu District, Shenzhen, where the time span is from January 1, 2015 to January 31, 2015 (*Zhu et al., 2022*). Weather data is divided into five categories: sunny, light rain, heavy rain, cloudy and foggy. POI data is divided into nine categories: business, transportation, medical, living, accommodation, education, food, shopping and others. Due to the limitations of experimental data collection, it is difficult to have existing

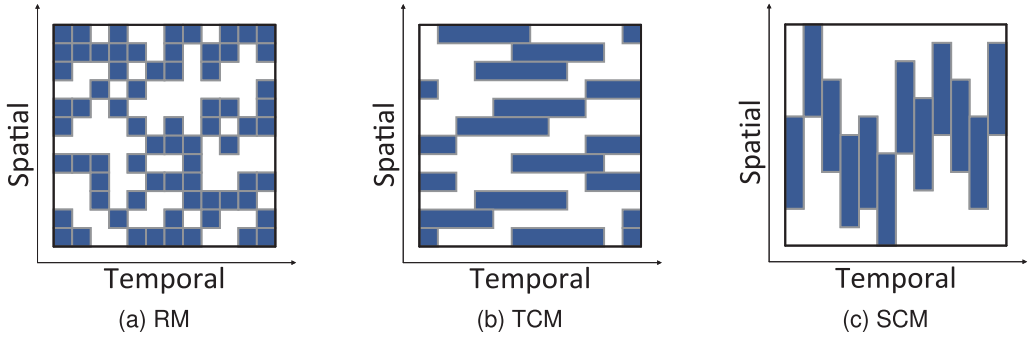

Figure 5 **Patterns of missing data.** Full-size DOI: 10.7717/peerj-cs.2408/fig-5

publicly available datasets that collect traffic speed data, road network data, and external correlates (Weather, POI, *etc*.,) at the same time, so only one regional dataset from Shenzhen is used in this experiment. Nevertheless, as our model is universal and transferable, researchers can validate it in any city that gives a relevant dataset.

Knowledge representation

The inputs to the knowledge representation model include attribute triple, road adjacency triple, and attribute co-occurrence triple. Based on the composition of the Shenzhen dataset, we construct POI attribute triples using road section number, POI category, and numbers of POI. We also use time, weather, and relevance to construct weather attribute triples, resulting in the construction of a knowledge graph of the Luohu District in Shenzhen. Specific examples are as follows: (road 123, enterprise, 3) indicates that there are three enterprises located around road section 123; (road 100, hospital, 1) indicates that a hospital exists around road section 100; (road 100, weather conditions, moment t) and (moment t, weather, clear) indicate that road 100 has clear weather condition at moment t.

Experimental design

In this article, 80% of the dataset is used as training data and the rest of the data was used as test data. We choose T = 288 time steps (*i.e*., 15 min $\times$ 288 = 72 h) as the imputation window. During training, we use the sliding window method to impute $[t, t + T]$, $[t + T, t + 2T]$, $[t + 2T, t + 3T]$, *etc*. We initialize all weight values uniformly and normalize the input data to [0,1]. Both GCN and BiLSTM_ATT networks contain two layers whose sizes are 128. The model is trained using the Adam optimizer with an initial learning rate of 0.01. As displayed in Fig. 5, we validate the performance of the model under three missing modes (*Li et al., 2018*; *Liang, Zhao & Sun, 2021*): 1) Random missing (RM), where missing values are completely independent of each other and displayed as randomly scattered points for each sensor (or road); 2) Temporal correlated missing (TCM), where missing values are dependent in the time dimension and appear as a consecutive time interval for each sensor (or road); 3) Spatially correlated missing (SCM), where missing values are dependent in the spatial dimension and appear at neighboring sensors or connected road links for each time slot. The performance of the model is compared with other baseline

Table 1 Embedded dimension selection effect value.

Dimensions	Missing patterns					
	RM		TCM		RCM	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
TransE(20)	0.0265	0.0485	0.0284	0.0477	0.0278	0.0509
TransE(50)	**0.0260**	0.0461	0.0275	0.0507	0.0274	0.0546
TransR(15)	0.0270	**0.0454**	**0.0264**	**0.0445**	**0.0265**	**0.0457**
TransR(20)	0.0370	0.0463	0.0265	0.0505	0.0325	0.0521
TransR(50)	0.0279	0.0613	0.0333	0.0521	0.0269	0.0556

Note:
Values in bold indicate the best result.

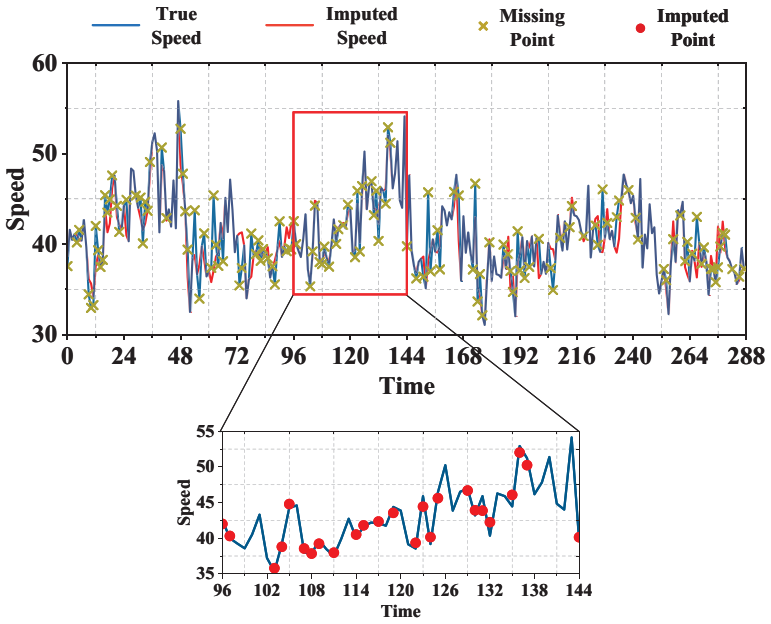

Figure 6 The visualization of data imputation results. Full-size DOI: 10.7717/peerj-cs.2408/fig-6

methods at different deletion rates from 30% to 80%. The model proposed in this study has an important hyperparameter, namely the knowledge embedding dimension, which has a significant impact on the data imputation results. After conducting multiple experiments, TransE and TransR models are used to learn the knowledge graph, and their performances are compared using embedding dimensions of 20 and 15 for TransE, and 15, 20, and 50 for TransR, as shown in Table 1. The comparison is performed in the typical scenario of 50% missing data, and the RMSE and MAE loss metrics are used to evaluate the models with different embedding dimensions. Based on the results, an embedding dimension of 15 is chosen to achieve the best final imputation results. Fig. 6 shows the traffic speed data imputation results under a missing rate of 50%, a knowledge embedding dimension of 15, a random missing pattern, and an RMSE evaluation metric.

Baselines

To demonstrate the effectiveness of our model in all aspects, we compared several baseline experiments. These include the machine learning methods: MEAN and SVR (*Wu, Ho & Lee, 2004*); the vector decomposition method: BGCP (*Chen, He & Sun, 2019*); the deep learning methods: GRU-D (*Che et al., 2018*), BRITS (*Cao et al., 2018*), GAIN (*Yoon, Jordon & Schaar, 2018*), PC-GAIN (*Wang et al., 2021*), GACN (*Ye, Zhang & Yu, 2021*), DGCRIN (*Kong et al., 2023*).

- **MEAN:** The missing elements are interpolated with the means of all relevant features.
- **SVR:** We choose support vector machines as the representative of regression-based machine learning imputation methods.
- **GRU-D:** GRU-D is a deep learning model architecture represented by two missing modes of masking and time interval, which improves model imputation performance through the application of the decay mechanism.
- **BRITS:** BRITS is a missing value imputation method for time series data based on RNN, which can directly learn missing values in a bidirectional recursive dynamical system without the need for any specific assumptions.
- **GAIN:** GAIN is a GAN-based method for unsupervised missing data imputation. It adds an indicator matrix to the GAN, which is to ensure that the generator generates samples according to the true underlying data distribution.
- **PC-GAIN:** PC-GAIN is a method of unsupervised missing data imputation. It proposes a kind of potential category information contained in a subset of low-missing rate data during pre-training while using synthetic pseudo-labels to identify auxiliary classifiers and then combines classifiers into GAN to help generators produce higher-quality prediction results.
- **BGCP:** BGCP extends the Bayesian probability matrix decomposition model to higher order tensor and applies it to the task of spatiotemporal traffic data. They focus not only on the configuration of the model, but also on the data representation (*i.e.*, matrix, third-order and fourth-order tensor).
- **GACN:** GACN is a graph attention convolutional network model for missing data imputation, which follows an encoder-decoder structure. As a typical spatiotemporal imputation model, GACN introduces a graph attention mechanism to learn the spatial correlation of adjacent sensors. In addition, it superimposes temporal convolutional layers to extract relationships in time series.
- **DGCRIN:** DGCRIN is also an imputation model based on dynamic graph convolutional networks and realizes state-of-the-art performance. It develops a graph generator to model dynamic spatial correlations and uses a dynamic graph convolutional gated recurrent unit to capture spatiotemporal relevances.

Evaluation metrics. We choose mean absolute error (MAE), root mean square error (RMSE) and mean absolute percentage error (MAPE) as evaluation indicators. Here's how the evaluation metrics are calculated:

Table 2 Experimental results of a data imputation method in RM pattern.

Models	Missing rate																	
	30%			40%			50%			60%			70%			80%		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
MEAN	0.0321	0.0868	2.16%	0.0476	0.1100	3.28%	0.0671	0.1361	5.16%	0.0911	0.1650	7.86%	0.1206	0.1973	9.45%	0.1540	0.2306	10.98%
SVR	0.0761	0.0880	5.67%	0.0783	0.0909	5.95%	0.0789	0.0921	6.13%	0.0831	0.0965	6.87%	0.0874	0.1035	7.03%	0.0951	0.1108	7.84%
GRU-D	0.0763	0.0944	5.88%	0.0822	0.1010	6.07%	0.0908	0.1113	7.16%	0.0917	0.1120	7.05%	0.1009	0.1222	8.20%	0.1044	0.1263	8.39%
BRITS	0.0526	0.0794	3.00%	0.0550	0.0818	3.73%	0.0580	0.0849	4.07%	0.0603	0.0867	4.54%	0.0624	0.0884	4.22%	0.0660	0.0923	4.51%
GAIN	0.0361	0.0547	2.07%	0.0399	0.0613	2.19%	0.0491	0.0734	2.84%	0.0617	0.1088	4.97%	0.0948	0.1716	8.14%	0.1173	0.1881	9.48%
PC-GAIN	0.0841	0.1417	6.85%	0.0875	0.1474	7.33%	0.0926	0.1524	8.21%	0.0955	0.1543	8.59%	0.1000	0.1582	9.13%	0.1080	0.1670	10.25%
BGCP	0.0573	0.0794	3.88%	0.0572	0.0794	3.97%	0.0574	0.0795	4.10%	0.0580	0.0801	4.54%	0.0584	0.0805	4.95%	0.0592	0.0817	5.05%
GACN	0.0534	0.0768	3.26%	0.0549	0.0771	3.39%	0.0557	0.0775	3.77%	0.0562	0.0782	3.56%	0.0575	0.0789	4.17%	0.0584	0.0800	4.35%
DGCRIN	0.0463	0.0711	1.75%	0.0469	0.0715	2.47%	0.0480	0.0723	2.61%	0.0503	0.0746	2.89%	0.0517	0.0757	3.59%	0.0523	0.0761	3.35%
KG-GAN	**0.0244**	**0.0427**	**1.44%**	**0.0254**	**0.0444**	**1.77%**	**0.0270**	**0.0454**	**2.22%**	**0.0291**	**0.0488**	**2.05%**	**0.0292**	**0.0523**	**2.58%**	**0.0309**	**0.0535**	**2.85%**

Note:
Values in bold indicate the best result.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} ||\widehat{y}_i - y_i|| \tag{14}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\widehat{y}_i - y_i)^2} \tag{15}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\% \tag{16}$$

where n represents the number of missing data, $\widehat{y}_i$ represents the prediction of missing value, and y_i represents the observed value.

Comparison experiment

Table 2 shows the error values of the KG-GAN and several baseline models under different missing rates and evaluation metrics. The data is the statistical result under RM missing pattern and we draw the following analyses: (1) Traditional imputation methods MEAN and SVR perform poorly compared to most deep learning methods, especially when the missing rate is higher than 50%; (2) compared to the traditional methods, RNN-based methods GRU-D and BRITS consider the time series correlation of data and have smaller imputation errors. However, the imputation performance of GRU-D is slightly lower than that of the BRITS algorithm, which is probably because GRU-D is more suitable for imputing medical data than traffic data; (3) GAIN and PC-GAIN are data imputation models based on data distribution. These GAN-based methods only adapt to low missing rate situations from the perspective of data distribution; (4) even at higher missing rates, the GACN model achieves a MAE of less than 6% and a RMSE of under 8%. Meanwhile, DGCRIN demonstrates even better performance. This highlights the significance of incorporating spatiotemporal correlations for traffic data imputation, which outperforms algorithms that only take temporal correlations and data distribution into account. (5) The

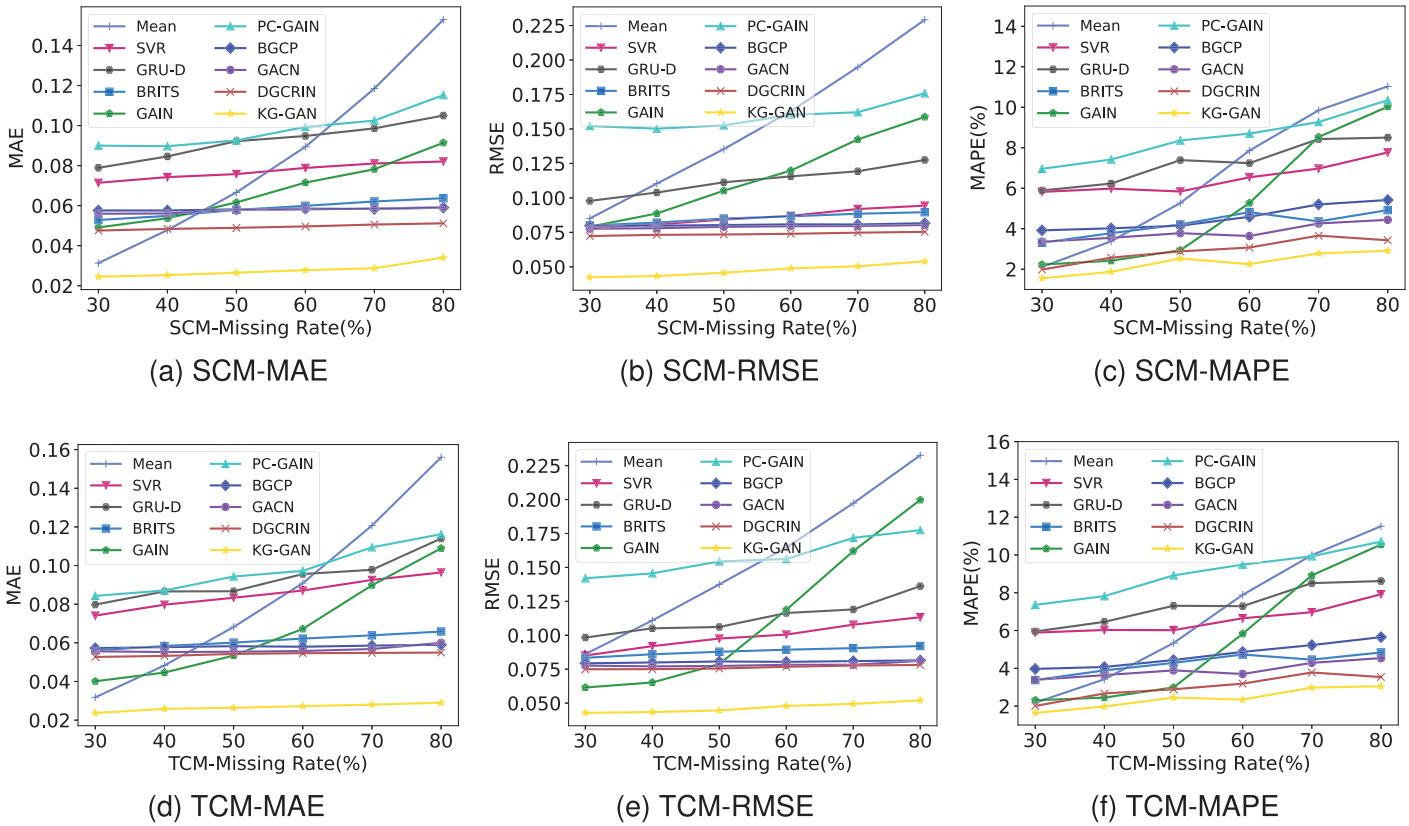

Figure 7 (A–F) Performance comparison in two missing patterns. Full-size DOI: 10.7717/peerj-cs.2408/fig-7

MAE, RMSE and MAPE of KG-GAN are lower than the best baseline DGCRIN and superior to other baselines, which fully illustrates the effectiveness of our model.

From the perspective of robustness analysis, Fig. 7 shows the performance comparison of various baseline models under SCM and TCM. The horizontal axis in the figure represents different data missing rates, and the vertical axis is used to visualize the loss results of different models under different evaluation metrics. Based on the experimental results, the following conclusions can be drawn: (1) The traditional MEAN method is basically unaffected by the missing pattern, and the loss value steadily increases with the increase of the missing rate in different evaluation metrics. When the missing rate is greater than 60%, its RMSE metric is higher than all other baseline models, indicating poor robustness; (2) the robustness of the SVR algorithm is superior to the MEAN method, and its performance under SCM is better than under TCM; (3) the GRU-D model mainly focuses on capturing time dependencies rather than spatial dependencies, so its performance under the time-continuous missing pattern is lower than under the space-continuous missing pattern. Moreover, as the missing rate increases, the loss value of GRU-D continues to rise, indicating that the robustness of the GRU-D model for different missing patterns and missing rates is not ideal; (4) intuitively, it can be found that the robustness of the GAIN model is only better than the traditional MEAN method. When

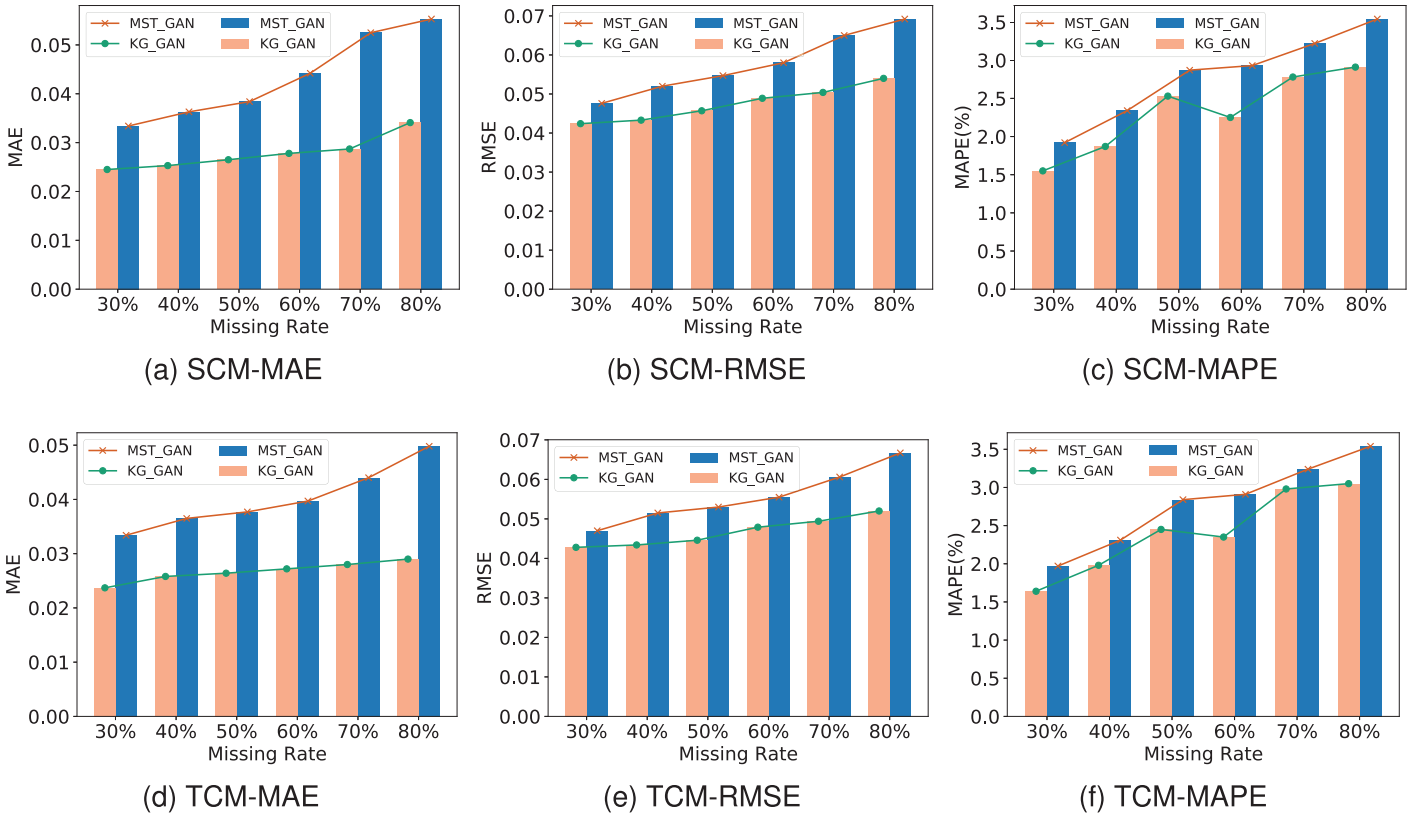

Figure 8 (A–F) Ablation experiment in SCM and TCM missing patterns. Full-size DOI: 10.7717/peerj-cs.2408/fig-8

the missing rate is greater than 40%, the loss curve of GAIN suddenly rises. This is because when the missing rate gradually increases, it is difficult for GAN to learn from historical data; (5) under different missing patterns and evaluation metrics, the robustness of BGCP, BRITS, GACN, DGCRIN and our KG-GAN are all very superior.

In general, since KG-GAN not only considers data distribution and spatiotemporal correlation dependencies but also introduces knowledge embedding of external related factors, the imputation performance of our model is the best.

Ablation experiments

To verify the gain effect of knowledge representation on the performance of generating adversarial network imputation data, this study compares KG-GAN with the MST-GAN model (*i.e.*, KG-GAN w/o EM-Cell), which differs in that KG-GAN introduces a knowledge learning module to model and process knowledge representation of multi-source traffic information. This study conducted ablation experiments under SCM and TCM in terms of MAE, RMSE and MAPE as shown in Fig. 8. The results intuitively suggest that as the rate of missing data increases, the accuracy of both models decreases. This is attributed to the fact that the increasing amount of missing data negatively impacts the models' learning ability. The study also indicates that in the case of missing data, it is

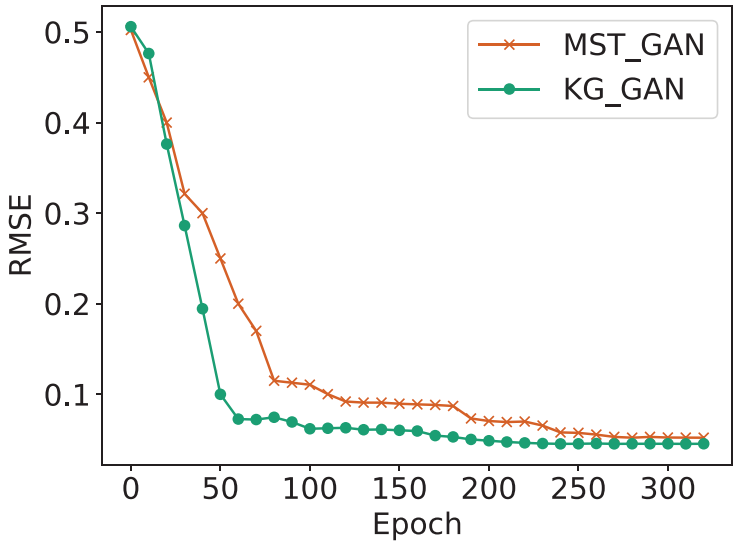

Figure 9 **Convergence rate in RM missing patterns.** Full-size DOI: 10.7717/peerj-cs.2408/fig-9

necessary to use models with better robustness to handle data, and KG-GAN's robustness is significantly better than MST-GAN's. In terms of model accuracy, KG-GAN's three indicators are better than MST-GAN's, especially in the MAE evaluation indicator, highlighting KG-GAN's superior imputation performance.

In addition, in order to verify that the convergence speed of the KG-GAN is better than that of the MST-GAN, an ablation experiment is also designed to measure the trend of the loss under RM missing pattern, with a data missing rate of 50% and a knowledge embedding dimension of 15, as shown in Fig. 9. The horizontal axis represents the training batch (epoch) of the model, and every 100 epochs takes an average of 5 s. One can see that the KG-GAN model converges faster than the MST-GAN model since the KG-GAN model learns the spatiotemporal features of the dataset faster during training, and can better capture the missing patterns in the data.

In summary, the results of the ablation experiments prove that external factors (such as weather, POI, *etc.*) and the complex correlations between external factors are of great significance to the imputation of traffic data. By using knowledge graphs as prior knowledge to guide the training of the model, the convergence speed, efficiency, and accuracy of the model can be improved, ultimately achieving high-quality imputation of traffic data.

CONCLUSION

In this article, we propose KG-GAN, a knowledge graph-enhanced model for spatiotemporal traffic data imputation, designed to improve deep learning-based imputation models' performance under high data scarcity and account for complex external factors. Our extensive experiments on a real-world traffic dataset validate KG-GAN's effectiveness in traffic data imputation. Specifically, we first construct an implicit knowledge representation of external factors using a fine-grained knowledge graph, which

accurately distinguishes between attributes and relationships. Next, we introduce a knowledge-aware embedding cell that integrates traffic data with the external knowledge representation, resulting in a refined traffic embedding. This embedding is then input into the MST-GAN model, facilitating effective convergence to the real traffic data distribution. Our approach achieves superior results by combining spatiotemporal feature learning with external knowledge, leading to more accurate imputation. The implications of our research are significant for traffic data imputation and intelligent transportation systems. By addressing data scarcity and incorporating external factors, our model offers a more robust solution for real-world applications, potentially improving traffic management and decision-making. Future work will focus on refining the knowledge graph to reduce noise, such as irrelevant connections between items and entities, which can impact model performance (*Yang et al., 2023*). We also plan to explore the application of KG-GAN to other domains and datasets to further validate its generalizability and robustness.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang 2023C01241, the National Natural Science Foundation of China under Grant 62072409 and Grant 62073295, and the Zhejiang Provincial Natural Science Foundation under Grant LR21F020003. There was no additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
"Pioneer" and "Leading Goose" R&D Program of Zhejiang: 2023C01241.
National Natural Science Foundation of China: 62072409 and 62073295.
Zhejiang Provincial Natural Science Foundation: LR21F020003.

Competing Interests

Xiangjie Kong is an Academic Editor for PeerJ. Zhenhui Xu and Junjie Zhou are employed by Zhejiang Supcon Information Co., Ltd.

Author Contributions

- Yinghui Liu conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Guojiang Shen conceived and designed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Nali Liu conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.

- Xiao Han conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Zhenhui Xu conceived and designed the experiments, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Junjie Zhou conceived and designed the experiments, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Xiangjie Kong conceived and designed the experiments, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The code and raw data are available in the Supplemental Files.

These data are from the original source at GitHub: https://github.com/lehaifeng/T-GCN/tree/master/KST-GCN.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.2408#supplemental-information.

REFERENCES

Bania RK, Halder A. 2020. R-ensembler: a greedy rough set based ensemble attribute selection algorithm with knn imputation for classification of medical data. *Computer Methods and Programs in Biomedicine* **184(4)**:105122 DOI 10.1016/j.cmpb.2019.105122.

Bordes A, Usunier N, Garcia-Durán A, Weston J, Yakhnenko O. 2013. Translating embeddings for modeling multi-relational data. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*. 2787–2795.

Caillault ÉP, Lefebvre A, Bigand A. 2020. Dynamic time warping-based imputation for univariate time series data. *Pattern Recognition Letters* **139(1)**:139–147 DOI 10.1016/j.patrec.2017.08.019.

Cao W, Wang D, Li J, Zhou H, Li L, Li Y. 2018. Brits: bidirectional recurrent imputation for time series. In: *Advances in Neural Information Processing Systems 31*.

Cetin M, Comert G. 2006. Short-term traffic flow prediction with regime switching models. *Transportation Research Record* **1965(1)**:23–31 DOI 10.1177/0361198106196500103.

Che Z, Purushotham S, Cho K, Sontag D, Liu Y. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports* **8**:6085 DOI 10.1038/s41598-018-24271-9.

Chen X, He Z, Sun L. 2019. A bayesian tensor decomposition approach for spatiotemporal traffic data imputation. *Transportation Research Part C: Emerging Technologies* **98(7)**:73–84 DOI 10.1016/j.trc.2018.11.003.

Chen X, Wang K, Li Z, Zhang Y, Ye Q. 2023. A novel nonconvex low-rank tensor completion approach for traffic sensor data recovery from incomplete measurements. *IEEE Transactions on Instrumentation and Measurement* **72**:1–15 DOI 10.1109/TIM.2023.3284929.

Chen X, Wang K, Zhao F, Deng F, Ye Q. 2024. Composite nonconvex low-rank tensor completion with joint structural regression for traffic sensor networks data recovery. *IEEE Transactions on Computational Social Systems* **11(5)**:6882–6896 DOI 10.1109/TCSS.2024.3406629.

Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. ArXiv DOI 10.48550/arXiv.1406.1078.

Gui H, Liu J, Tao F, Jiang M, Norick B, Kaplan LM, Han J. 2017. Embedding learning with events in heterogeneous information networks. *IEEE Transactions on Knowledge and Data Engineering* **29(11)**:2428–2441 DOI 10.1109/TKDE.2017.2733530.

Jianrui X, Xingyi L, Yi H. 2010. Short-term traffic flow forecasting model under missing data. *Journal of Computer Applications* **30(4)**:1117–1120 DOI 10.3724/SP.J.1087.2010.01117.

Kaiser J. 2014. Dealing with missing values in data. *Journal of Systems Integration* **5(1)**:42–51.

Kong X, Shen Z, Wang K, Shen G, Fu Y. 2024. Exploring bus stop mobility pattern: a multi-pattern deep learning prediction framework. *IEEE Transactions on Intelligent Transportation Systems* **25(7)**:6604–6616 DOI 10.1109/TITS.2023.3345872.

Kong X, Zhou W, Shen G, Zhang W, Liu N, Yang Y. 2023. Dynamic graph convolutional recurrent imputation network for spatiotemporal traffic missing data. *Knowledge-Based Systems* **261(3)**:110188 DOI 10.1016/j.knosys.2022.110188.

Lana I, Del Ser J, Velez M, Vlahogianni EI. 2018. Road traffic forecasting: recent advances and new challenges. *IEEE Intelligent Transportation Systems Magazine* **10(2)**:93–109 DOI 10.1109/MITS.2018.2806634.

Li X, Ma J, Yu J, Zhao M, Yu M, Liu H, Ding W, Yu R. 2023. A structure-enhanced generative adversarial network for knowledge graph zero-shot relational learning. *Information Sciences* **629**:169–183 DOI 10.1016/j.ins.2023.01.113.

Li L, Zhang J, Wang Y, Ran B. 2018. Missing value imputation for traffic-related time series data based on a multi-view learning method. *IEEE Transactions on Intelligent Transportation Systems* **20(8)**:2933–2943 DOI 10.1109/TITS.2018.2869768.

Liang Y, Zhao Z, Sun L. 2021. Dynamic spatiotemporal graph convolutional neural networks for traffic data imputation with complex missing patterns. ArXiv DOI 10.48550/arXiv.2109.08357.

Lin Y, Liu Z, Sun M. 2016. Knowledge representation learning with entities, attributes and relations. In: *International Joint Conference on Artificial Intelligence*.

Lin Y, Liu Z, Sun M, Liu Y, Zhu X. 2015. Learning entity and relation embeddings for knowledge graph completion. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. Washington, D.C.: AAAI.

Liu Y, Shen G, Cui C, Zhao Z, Han X, Du J, Zhao X, Kong X. 2024. Kddc: knowledge-driven disentangled causal metric learning for pre-travel out-of-town recommendation. In: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*. 2207–2215.

Muppalla R, Lalithsena S, Banerjee T, Sheth A. 2017. A knowledge graph framework for detecting traffic events using stationary cameras. In: *Proceedings of the 2017 ACM on Web Science Conference*, 431–436.

Ni Q, Cao X. 2022. Mbgan: an improved generative adversarial network with multi-head self-attention and bidirectional rnn for time series imputation. *Engineering Applications of Artificial Intelligence* **115(1)**:105232 DOI 10.1016/j.engappai.2022.105232.

Ning Y, Liu H, Wang H, Zeng Z, Xiong H. 2024. Uukg: unified urban knowledge graph dataset for urban spatiotemporal prediction. In: *Advances in Neural Information Processing Systems 36*.

Olayode IO, Du B, Tartibu LK, Alex FJ. 2024. Traffic flow modelling of long and short trucks using a hybrid artificial neural network optimized by particle swarm optimization. *International Journal of Transportation Science and Technology* **14**:137–155 DOI 10.1016/j.ijtst.2023.04.004.

Pamuła T. 2018. Impact of data loss for prediction of traffic flow on an urban road using neural networks. *IEEE Transactions on Intelligent Transportation Systems* **20(3)**:1000–1009 DOI 10.1109/TITS.2018.2836141.

Peng C, Xia F, Naseriparsa M, Osborne F. 2023. Knowledge graphs: opportunities and challenges. *Artificial Intelligence Review* **56(11)**:13071–13102 DOI 10.1007/s10462-023-10465-9.

Qu L, Li L, Zhang Y, Hu J. 2009. Ppca-based missing data imputation for traffic flow volume: a systematical approach. *IEEE Transactions on Intelligent Transportation Systems* **10(3)**:512–522 DOI 10.1109/TITS.2009.2026312.

Shen G, Liu N, Liu Y, Zhou W, Kong X. 2022. Traffic flow imputation based on multi-perspective spatiotemporal generative adversarial networks. In: *Proceedings of CECNet 2022*. Amsterdam: IOS Press, 62–73.

Shen G, Zhou W, Zhang W, Liu N, Liu Z, Kong X. 2023. Bidirectional spatial–temporal traffic data imputation via graph attention recurrent neural network. *Neurocomputing* **531(21)**:151–162 DOI 10.1016/j.neucom.2023.02.017.

Tan H, Wu Y, Cheng B, Wang W, Ran B. 2014. Robust missing traffic flow imputation considering nonnegativity and road capacity. *Mathematical Problems in Engineering* **2014(2)**:1–8 DOI 10.1155/2014/763469.

Tan Q, Ye M, Yang B, Liu S, Ma AJ, Yip TC-F, Wong GL-H, Yuen P. 2020. Data-gru: dual-attention time-aware gated recurrent unit for irregular multivariate time series. *Proceedings of the AAAI Conference on Artificial Intelligence* **34(1)**:930–937 DOI 10.1609/aaai.v34i01.5440.

Tian Y, Zhang K, Li J, Lin X, Yang B. 2018. Lstm-based traffic flow prediction with missing data. *Neurocomputing* **318(5)**:297–305 DOI 10.1016/j.neucom.2018.08.067.

Wang Y, Li D, Li X, Yang M. 2021. PC-GAIN: pseudo-label conditional generative adversarial imputation networks for incomplete data. *Neural Networks* **141(1)**:395–403 DOI 10.1016/j.neunet.2021.05.033.

Wang Z, Zhang J, Feng J, Chen Z. 2014. Knowledge graph embedding by translating on hyperplanes. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 28.

Wu C, Ho J, Lee D. 2004. Travel-time prediction with support vector regression. *IEEE Transactions on Intelligent Transportation Systems* **5(4)**:276–281 DOI 10.1109/TITS.2004.837813.

Xu Z, Lv Z, Li J, Sun H, Sheng Z. 2022. A novel perspective on travel demand prediction considering natural environmental and socioeconomic factors. *IEEE Intelligent Transportation Systems Magazine* **15(1)**:136–159 DOI 10.1109/MITS.2022.3162901.

Xu Z, Zhang H, Hu C, Mei L, Xuan J, Choo K-KR, Sugumaran V, Zhu Y. 2016. Building knowledge base of urban emergency events based on crowdsourcing of social media. *Concurrency and Computation: Practice and Experience* **28(15)**:4038–4052 DOI 10.1002/cpe.3780.

Xu T, Zhang P, Huang Q, Zhang H, Gan Z, Huang X, He X. 2018. Attngan: fine-grained text to image generation with attentional generative adversarial networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 1316–1324.

Yang Y, Huang C, Xia L, Huang C. 2023. Knowledge graph self-supervised rationalization for recommendation. In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. New York: ACM, 3046–3056.

Ye Y, Zhang S, Yu JJ. 2021. Spatial-temporal traffic data imputation via graph attention convolutional network. In: Farkaš I, Masulli P, Otte S, Wermter S, eds. *Artificial Neural Networks and Machine Learning—ICANN 2021. ICANN 2021. Lecture Notes in Computer Science*. Vol. 12891. Cham: Springer, 241–252 DOI 10.1007/978-3-030-86362-3_20.

Yi X, Walia E, Babyn P. 2019. Generative adversarial network in medical imaging: a review. *Medical Image Analysis* **58(2)**:101552 DOI 10.1016/j.media.2019.101552.

Yoon J, Jordon J, Schaar M. 2018. Gain: missing data imputation using generative adversarial nets. In: *International Conference on Machine Learning*. 5689–5698.

Zhang W, Zhang P, Yu Y, Li X, Biancardo SA, Zhang J. 2021. Missing data repairs for traffic flow with self-attention generative adversarial imputation net. *IEEE Transactions on Intelligent Transportation Systems* **23(7)**:7919–7930 DOI 10.1109/TITS.2021.3074564.

Zhu J, Han X, Deng H, Tao C, Zhao L, Wang P, Lin T, Li H. 2022. Kst-gcn: a knowledge-driven spatial-temporal graph convolutional network for traffic forecasting. *IEEE Transactions on Intelligent Transportation Systems* **23(9)**:15055–15065 DOI 10.1109/TITS.2021.3136287.