BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 05/25 pp621–636
DOI: 10.26599/BDMA.2024.9020020
Volume 7, Number 3, September 2024

A Large-Scale Spatio-Temporal Multimodal Fusion Framework for Traffic Prediction

Bodong Zhou, Jiahui Liu, Songyi Cui, and Yaping Zhao*

Abstract: Traffic prediction is crucial for urban planning and transportation management, and deep learning techniques have emerged as effective tools for this task. While previous works have made advancements, they often overlook comprehensive analyses of spatio-temporal distributions and the integration of multimodal representations. Our research addresses these limitations by proposing a large-scale spatio-temporal multimodal fusion framework that enables accurate predictions based on location queries and seamlessly integrates various data sources. Specifically, we utilize Convolutional Neural Networks (CNNs) for spatial information processing and a combination of Recurrent Neural Networks (RNNs) for final spatio-temporal traffic prediction. This framework not only effectively reveals its ability to integrate various modal data in the spatio-temporal hyperspace, but has also been successfully implemented in a real-world large-scale map, showcasing its practical importance in tackling urban traffic challenges. The findings presented in this work contribute to the advancement of traffic prediction methods, offering valuable insights for further research and application in addressing real-world transportation challenges.

Key words: spatio-temporal; traffic prediction; multimodal fusion; learning representation

1 Introduction

In recent decades, the automobile industry has experienced a remarkable growth, leading to a noticeable rise in both the production and ownership of vehicles. According to the annual report of the China Association of Automobile Manufacturers, new energy vehicle production in 2022 increased by 90.5% year-on-year. The total number of motor vehicles in China reached 430 million, and the number of licensed drivers reached 520 million in 2023[1]. The increase in private vehicle holdings has been accompanied by rapid urbanization process, resulting in a mounting need for efficient mobility networks within urban areas, which have led to higher traffic flows on road networks. The intensifying road traffic has contributed to more severe traffic congestion issues in urban areas, calling for adopting traffic speed prediction to develop effective traffic flow management approaches[2, 3]. With reliable predictions of future traffic speeds and volumes, authorities can now optimize traffic signal timing and provide routing recommendations to dynamically alleviate bottlenecks. Strategies like flexible signal coordination and dynamic route guidance can be employed, helping to reduce travel times and emissions[4]. Furthermore, traffic speed prediction plays a vital role in Intelligent Transportation Systems (ITS), smart city initiatives,

• Bodong Zhou is with Technical Consulting Department, Shanghai EchoBlend Internet Technology Co. Ltd., Shanghai 201111, China. E-mail: luckkyzhou@gmail.com.
• Jiahui Liu and Yaping Zhao are with Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong 999077, China. E-mail: jliu1@connect.hku.hk; zhaoyp@connect.hku.hk.
• Songyi Cui is with Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong, Hong Kong 999077, China. E-mail: echo07@connect.hku.hk.
* To whom correspondence should be addressed.

Manuscript received: 2023-10-16; revised: 2024-02-23; accepted: 2024-03-22

and intelligent driving systems[5–13]. By leveraging real-time traffic data, cities can optimize transportation infrastructure, improve public transportation services, and reduce environmental impacts. Consequently, the domain of traffic prediction has been receiving escalating scholarly attention, as evidenced by the burgeoning corpus of literature on the subject[14–19].

Previous researchers preferred to apply classical statistical models to predict traffic volumes[20, 21], of which the Auto Regressive Integrated Moving Average (ARIMA) family of models is the most widely used[22–24]. At that time, the data sources rely on historical data, such as loop detector data or GPS traces[25–27]. However, these methods suffer from limitations, such as insufficient service coverage, limited temporal resolution, and lack of real-time updates. The emergence of technologies like smartphones and connected vehicles has provided a wealth of real-time data that can be leveraged to improve the accuracy of traffic speed prediction[28–34]. The diversification of data and advancements in computer technology have made more accurate traffic prediction possible. Therefore, recent studies have predominantly employed machine learning methodologies, such as the hidden markov model, fuzzy neural network, and support vector machines, to enhance the models' ability to handle high-dimensional data and capture nonlinear relationships in traffic prediction problems[15, 16, 35].

In terms of spatial modeling, many approaches rely on static graphs for graph convolution[36]. When it comes to temporal modeling, the majority of existing strategies fall into one of three categories: those based on Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and attention-based approaches[37–41]. While previous forecasting methods primarily rely on temporal and spatial information, and some studies have considered spatio-temporal information, they have not taken into account the application of multimodal data. Although these methods have achieved some success, they have not considered temporal-spatial heterogeneity, limiting their ability to capture complex relationships between different locations and their impact on traffic speeds. Furthermore, these methods have been evaluated on small-scale datasets, which may not generalize well to real-world scenarios.

Thus, developing models which are capable of real-time network-wide traffic speed prediction poses tough challenges by utilizing comprehensive spatio-temporal traffic data. One of the critical challenges is how to effectively extract the complex spatio-temporal correlations between different nodes. Because observations from neighboring locations and timestamps are not independent, but rather dynamically interrelated[42, 43]. For example, in the context of point-of-interest recommendation, as information spreads online, the popularity of a certain place at a certain moment will strongly affect the traffic conditions near this place. Models aimed at forecasting traffic conditions need to model how conditions propagate from one area to adjacent regions or cascade forward in time according to different data sources. Prediction approaches can reach their full potential only by adequately representing these spatio-temporal interaction patterns. Therefore, how to mine the non-linear and complex spatio-temporal data to discover the underlying spatio-temporal patterns and make accurate traffic flow predictions is a highly challenging problem. Recently, Zhou et al.[44] proposed a method that leverages spatial information from large-scale map data to predict traffic congestion. While their approach demonstrates the potential of using spatial information alone, it still has limitations in terms of prediction accuracy and does not incorporate temporal information from historical traffic speeds. Overall, existing researches show that few studies can predict network-wide traffic speed propagation dynamics with multimodal data source, signifying a significant research gap and calls for innovative solutions.

To cope with the aforementioned issues, we present a model to solve the problem of traffic speed prediction. Previous methods fail to capture the multimodal data in different layers. To solve this, our method integrates spatial and temporal data, providing a comprehensive synthesis of the abundant information found on expansive maps, which enhances the accuracy in predicting traffic scenarios. As a result, the diverse global multimodal data pertinent to traffic forecasting is consolidated into a spatio-temporal hyperplane[45]. This hyperplane can then be effectively utilized in conjunction with mapped locations to forecast traffic situations at various points in both space and time. In order to better fuse the rich information on a large-scale map and infer traffic situations, a novel and efficient framework is proposed in this paper, so that the global multimodal information referenced by traffic prediction is aggregated into a geo-preserving representation in a high-dimensional superspace, which can be utilized with mapped locations to predict traffic situations at different locations and time points (see Fig. 1).

The main contributions of this work are summarized as follows:

- We propose a novel approach for large-scale real-time network-wide traffic speed prediction by integrating various modalities from both spatial and temporal aspects. Our approach successfully fuses these rich information which includes social media texts, real estates, points of interest, and traffic speed, enabling its practical application in real-world scenarios. This distinguishes it from previous work[44] that only focuses on the spatial aspect.
- We introduce an innovative framework, large-scale spatio-temporal multimodal fusion framework, which effectively captures the complex relationships between spatial and temporal features in traffic data. This complex framework leverages advanced techniques, such as recurrent neural networks and fusion encoders, to enhance the accuracy and robustness of the prediction.
- Extensive experiments have been conducted on a real large-scale map dataset to evaluate the performance of the approach. Our approach achieves improved prediction accuracy compared to existing methods that only consider the fusion of map spatial information. This demonstrates the effectiveness and practicality of our approach in real-world traffic planning and management.

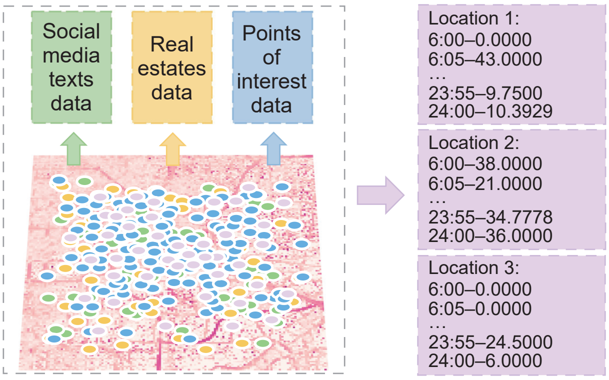


Fig. 1 A variety of multimodal information is created at different locations on a large-scale map of the real world. Traffic speed prediction at any query location can be learned based on these diverse information.

2 Related Work

2.1 Spatial traffic prediction

The field of traffic forecasting has been extensively studied for several decades, primarily in civil engineering and traffic engineering research. Over time, researchers have made advancements in understanding traffic patterns and developing sophisticated forecasting models. Existing studies show that traffic speed prediction has primarily been tackled using three approaches: statistical methods, traditional machine learning methods, and deep learning methods[19]. In recent years, there has been a shift towards using deep learning techniques for traffic prediction[46–48]. Among them, CNNs have been widely used[49–52]. These models map traffic speed values onto the geographical space and leverage CNNs to extract meaningful spatial features and patterns from the data, leading to accurate predictions of future traffic conditions. The power of deep learning allows the CNN models to effectively capture and analyze the complex spatial relationships in traffic data, resulting in improved prediction accuracy compared to traditional methods. The utilization of CNNs in traffic prediction demonstrates the potential of deep learning techniques in addressing real-world transportation challenges.

Li et al.[46] employed a Diffusion Convolutional Recurrent Neural Network (DCRNN) to predict traffic flow, modeling the traffic flow as a directed graph. They captured the spatial dependencies of traffic flow by utilizing bidirectional random walks on the graph. Wang et al.[47] developed a graph-based neural network approach that utilizes a deep learning framework to capture the topological structure of road networks. They employed a regression Graph Recursive Neural Network (GRNN) to track speed variations across the network, using trajectory data to predict average traffic speed. Zhang et al.[48] presents a Gated Recurrent Units (GRU) based multitask deep learning model for predicting network-wide traffic speed, enhanced by two performance-improving methods: a nonlinear Granger causality test and Bayesian optimization. Nguyen et al.[49] designed an EO-CNN model for traffic prediction, which is based on the Equilibrium Optimizer (EO), a metaheuristic algorithm. Mehdi et al.[51] proposed an approach based on differential entropy for labeling congestion levels and developed a supervised congestion prediction model using a CNN. The study incorporates various traffic meta-parameters, such as node localization, date, day of the week, time of day, special road conditions, and holidays. The model is validated on the CityPulse dataset, which contains vehicle traffic records collected in Aarhus city, Denmark, over a six-month period. The simulation results demonstrate accurate predictions of congestion rates for different observation nodes. The authors suggest that this system can help prevent traffic congestion by redirecting drivers to alternative routes.

Furthermore, Zhou et al.[44] proposed an end-to-end framework based on CNNs that learns geo-preserving representations from a real-world large-scale dataset to predict traffic congestion situations. The framework effectively predicts traffic congestion at any location on a large-scale map by aggregating comprehensive congestion factors and leveraging global reference information. The framework incorporates a multimodal fusion module and a representation mapping module to achieve accurate predictions of traffic congestion.

2.2 Spatio-temporal traffic prediction

Predicting traffic solely based on spatial information is insufficient since traffic data exhibit temporal characteristics. To improve the accuracy of traffic prediction by leveraging the synergy between spatial and temporal information, several approaches have been proposed that combine RNNs with either K-Nearest Neighbors (KNNs) or CNNs[53–57].

Min and Wynter[53] developed an approach based on extended time series, utilizing a multivariate spatiotemporal autoregression (namely MSTAR) model to interpret transient behaviors on traffic networks. Yu et al.[54] introduced a network grid representation method that preserves the detailed structure of transportation networks. They convert network-wide traffic speeds into static images and input them into a novel deep architecture called Spatiotemporal Recurrent Convolutional Networks (SRCNs) for traffic forecasting. Qiu et al.[55] employed various recurrent neural network architectures, adopting a multi-task learning approach to explore spatial and temporal correlations between different cellular regions. In pursuit of enhanced predictive accuracy, Luo et al.[56] proposed a spatiotemporal traffic flow prediction method that combines KNN with Long Short-Term Memory (LSTM) networks. In this approach, KNN is employed to capture spatial features, while LSTM is utilized to capture temporal features. This model is referred to as the KNN-LSTM model. Wang et al.[57] developed a model using Bidirectional Long Short-Term Memory Neural Networks (Bi-LSTM NNs) to represent the structure of critical road networks. They then fed the captured spatiotemporal features into a fully connected layer for predicting network traffic speed. An empirical study is conducted to demonstrate the interpretability of their model.

Pan et al.[39] proposed a deep-meta-learning based model called ST-MetaNet, which collectively predicts traffic for all locations simultaneously. ST-MetaNet utilizes a sequence-to-sequence architecture with an encoder to learn historical information and a decoder to make predictions step by step. The encoder and decoder have the same network structure, consisting of a recurrent neural network to encode the traffic, a meta graph attention network to capture spatial correlations, and a meta recurrent neural network to consider temporal correlations. This structure successfully addresses the challenges of complex spatio-temporal correlations in urban traffic and the diversity of such correlations across locations.

In summary, while the review discusses the utilization of various traffic meta-parameters and trajectory data, it falls short of addressing the integration of heterogeneous data sources, such as social media, weather conditions, or unexpected events, which could potentially improve the accuracy of predictions. Moreover, the review does not tackle the scalability of these models to more extensive and intricate road networks, nor does it consider their computational efficiency—both critical aspects for practical applications in the real world.

In our research, we conduct a comprehensive analysis of spatio-temporal distributions using three distinct spatial sources: text, other inputs, and traffic speed as temporal features. To accomplish this, we employ convolution operations on these three modalities. By combining sentences and numerical values in a 2D map space, we successfully learn a joint representation. Additionally, we utilize a recurrent structure to effectively capture the interdependencies between these joint features and speed inputs. This enables accurate prediction of specific traffic speed values based on location queries. As a result, the distribution of location queries and multimodal representations are seamlessly integrated into a unified mapping space.

3 Methodology

3.1 Overview

Previous traffic prediction models have focused only on spatial or temporal information and have used relatively simple neural network structures, resulting in unsatisfactory prediction accuracy. To address issues above, we propose a spatio-temporal multimodal end-to-end traffic speed prediction network. This approach utilizes a CNN to learn spatial information and a GRU network to predict the final speed sequence values at any location on a large-scale map efficiently. It also incorporates multimodal social data with the location of the large-scale map as global reference information.

The global reference information is divided into three categories: social media texts, real estates, and points of interest. These categories are scattered across the entire map with different distributions (see Fig. 1). The multimodal global reference information is formatted in a learnable and adaptive manner and is learned and fused into a global-aware geo-preserving representation, also known as a meta-representation. This meta-representation aggregates all reference information from the entire large-scale map. Additionally, the geo-temporal representation of speed values is learned as a crucial part of the spatio-temporal fusion structure. A geo-preserving representation is also learned to match a specific location, which is used to capture traffic speed situations. Finally, the global-aware representation, temporal-aware representation, and location-aware representation are aggregated and learned to obtain location-specific traffic speed results.

By incorporating both spatial and temporal information, as well as multimodal global reference information, our proposed model aims to improve the accuracy of traffic speed prediction. The use of advanced neural network structures and the fusion of various data sources contribute to a more detailed and comprehensive analysis of traffic patterns.

3.2 Problem statement

The task of predicting traffic speed can be effectively addressed using the proposed method, yielding valuable results. The first step is to grid the entire large-scale map in order to characterize the data. This results in the original map being formatted as a 2-dimensional grid array with dimensions of $H \times W$. Each grid on the map is associated with three types of global reference information in the region, which are quantized as multi-dimensional features. These features are then placed into the corresponding location on the gridded map, resulting in three multi-channel matrices.

Specifically, for each grid, a vector of length D_{mt} is generated to represent the social media information from the dataset after preprocessing. If a grid does not have any social media information, a zero-vector is used instead. Similarly, the preprocessed real estate information and points of interest information for each grid can be represented by vectors of length D_{re} and D_{pi}, respectively. Consequently, the social media texts, real estate data, and points of interest information for the entire gridded map can be formulated as three multi-channel matrices: F_{mt}, F_{re}, and F_{pi}, respectively. These matrices have dimensions of $H \times W \times D_{\mathrm{mt}}$, $H \times W \times D_{\mathrm{re}}$, and $H \times W \times D_{\mathrm{pi}}$, respectively, are utilized as part of the input for the proposed framework.

For the gridded map with dimensions $H \times W$, different regions are divided into training, validation, and test regions. The data within these regions are then aggregated into training, validation, and test sets. For the entire gridded map, each grid $x_{h,w}$ is indexed by its location in the h-th row and w-th column, where $h = 1, 2, \ldots, H$, and $w = 1, 2, \ldots, W$. For each $x_{h,w}$, the proposed model is capable of generating results $\hat{y}^{L'}_{h,w}$, which is a vector of length L' representing multiple time points in a day. Given a query point, the length of the speed sequence $V^{L}_{h,w}$ can be learned as a temporal-aware representation, where L denotes additional time points preceding L'. By incorporating the global-aware reference information and the temporal-aware representation, the task is to predict the speed values of roads at subsequent time points after the given one at any unvisited grid area (test area). The objective of the proposed method is to minimize the difference between the predicted $\hat{y}^{L'}_{h,w}$ and the ground truth values $y^{L'}_{h,w}$ for the corresponding location.

3.2.1 Framework and methods

The traffic speed prediction task is addressed by a novel, spatio-temporal, and lightweight framework, that effectively learns and generalizes global reference information and temporal reference information. This enables the framework to handle predictions by matching and fusing specific query points.

To achieve this, the framework, as show in Fig. 2,

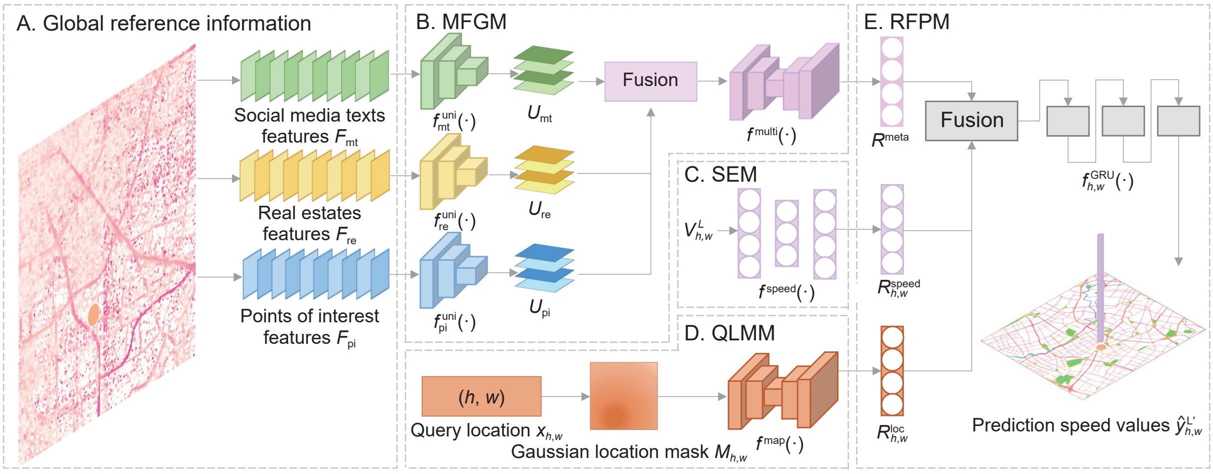

A. Global reference information
Social media texts features $F_{\rm mt}$
Real estates features $F_{\rm re}$
Points of interest features $F_{\rm pi}$
B. MFGM
$f_{\rm mt}^{\rm uni}(\cdot)$
$U_{\rm mt}$
$f_{\rm re}^{\rm uni}(\cdot)$
$U_{\rm re}$
$f_{\rm pi}^{\rm uni}(\cdot)$
$U_{\rm pi}$
Fusion
$f^{\rm multi}(\cdot)$
C. SEM
$V_{h,w}^{L}$
$f^{\rm speed}(\cdot)$
D. QLMM
(*h*, *w*)
Query location $x_{h,w}$
Gaussian location mask $M_{h,w}$
$f^{\rm map}(\cdot)$
E. RFPM
$R^{\rm meta}$
$R_{h,w}^{\rm speed}$
$R_{h,w}^{\rm loc}$
Fusion
$f_{h,w}^{\rm GRU}(\cdot)$
Prediction speed values $\hat{y}_{h,w}^{L'}$

Fig. 2 The proposed framework. Firstly, the featured global reference information is fed into the MFGM to obtain a global-aware meta-representation in a high-dimensional superspace. In addition, speed value at the query point is fed into the SEM to obtain the temporal-aware speed-representation. Also, the QLMM maps a query location as a location-aware mapped-representation in the same superspace. Finally, the RFPM makes prediction on traffic speed based on the global-aware meta-representation, temporal-aware speed-representation, and location-aware mapped-representation.

utilizes featured global reference information ($F_{\rm mt}$, $F_{\rm re}$, and $F_{\rm pi}$) of the entire large-scale map. These features are first inputted into the Multimodal Fusion and Generalization Module (MFGM) (Part B in Fig. 2) to obtain the global-aware meta-representation of the map. Simultaneously, based on a query location ($x_{h,w}$), the Speed Encoder Module (SEM) (Part C in Fig. 2) learns the temporal-aware speed-representation. Additionally, the query location is transformed into a location-aware mapped-representation through the Query Location Mapping Module (QLMM) (Part D in Fig. 2). Finally, the Representation Fusion and Prediction Module (RFPM) (Part E in Fig. 2) combines the global-aware meta-representation, temporal-aware speed-representation, and location-aware mapped-representation to make predictions using a neural network.

To be specific, MFGM is tailored to extract three distinct global map features, employing three separate convolutional layers accompanied by corresponding linear layers. Subsequently, these three linear layers undergo concatenation before being fed into another convolutional network to extract the fusion linear features of the map. Meanwhile, SEM employs multiple linear layers to extract speed features. QLMM initially expands the input location points into two-dimensional global features using a Gaussian distribution and then utilizes convolutional networks and linear layers for feature extraction. RFPM consolidates the linear features obtained from the previous three components through concatenation, feeding them into a GRU network[58] to ultimately predict the speed values corresponding to the current position on the map.

During training, all features of the global reference information are consistently fed into the MFGM at each step. This ensures that the MFGM is robust enough to comprehensively extract and generalize multimodal global reference information, resulting in the global-aware meta-representation. Consequently, during inference, the stable and informative global-aware meta-representation, which is saved after training, is directly inputted into the RFPM without the need for the heavy MFGM architecture and complex global reference information. This design choice allows the framework to maintain lightweight parameters and achieve incredible efficiency.

The details of each module in the proposed framework will be introduced below.

3.2.2 MFGM

Different kinds of global reference information are formatted into a matrix with various channels. To obtain three unimodal representations of the same size $H_{\rm ur}\times W_{\rm ur}\times D_{\rm ur}$, where $H_{\rm ur}$, $W_{\rm ur}$, and $D_{\rm ur}$ denote the dimensions of the output of neural networks. Three distinct convolutional neural networks are leveraged, denoted as $\left\{f_m^{\rm uni}(\cdot)\right\}_{m={\rm mt,\,re,\,pi}}$, where each represents convolutional neural network which is designed for $F_{\rm mt}$, $F_{\rm re}$, and $F_{\rm pi}$. These networks aim to capture the unique characteristics of each type of global reference information for F_{mt}, F_{re}, and F_{pi}, denoted as $\{U_m\}_{m=\mathrm{mt,\,re,\,pi}}$, where any $U_m \in \mathrm{R}^{H_{\mathrm{ur}} \times W_{\mathrm{ur}} \times D_{\mathrm{ur}}}$, and the forward procession can be formulated as

$$U_m = f_m^{\mathrm{uni}}(F_m), \quad m = \mathrm{mt,\ re,\ pi} \qquad (1)$$

To create a joint representation, the three unimodal representations are channel-wise concatenated. This representation is then fed into a convolutional neural network, $f^{\mathrm{multi}}(\cdot)$, which learns joint representation of multiple features, similar to an auto-encoder. The purpose of this network is to mix the representations from different modalities and shape them into a global reference-rich multi-channel representation. Importantly, the resulting representation maintains the spatial structure of the original map on each channel.

This final representation is referred to as the global-aware meta-representation. It encapsulates the combined information from the different types of global reference information and serves as a comprehensive and detailed representation of the data. This global-aware meta-representation is denoted as R^{meta}, where R^{meta} is a one-dimensional neural vector, which is obtained by

$$R^{\mathrm{meta}} = f^{\mathrm{multi}}(\mathrm{concat}\,[U_{\mathrm{mt}}, U_{\mathrm{re}}, U_{\mathrm{pi}}]) \qquad (2)$$

3.2.3 SEM

For each query location $x_{h,w}$, it is necessary to utilize the temporal continuity information of traffic speeds to predict the subsequent speed values. Therefore, the first step is to transform the speed values within a certain time period into neural representations using a MultiLayer Perceptron (MLP) network. This conversion allows us to obtain one-dimensional vector representations for speed $R_{h,w}^{\mathrm{speed}}$ through the fusion encoder, which combines these speed representations with other spatial representations on the map.

To be more specific, the original speed values $V_{h,w}^{L}$, which monitored time series' length is L, and its each single element is a time point among that series. Those would be input into the speed encoder, denoted as $f^{\mathrm{speed}}(\cdot)$ at the current query point (h, w), resulting in the generation of dense representations for speed $R_{h,w}^{\mathrm{speed}}$ for each $x_{h,w}$. The forward process is formulated as

$$R_{h,w}^{\mathrm{speed}} = f^{\mathrm{speed}}(V_{h,w}^{L}),\ h \in [1, H], w \in [1, W] \qquad (3)$$

3.2.4 QLMM

Describing the spatial relationship between each query location in a gridded map and the entire map using coordinate pairs $(x_{h,w})$ is challenging. Instead, a more effective approach is to generate a learnable geo-preserving representation that encapsulates the location information. This representation takes the form of a location-rich multi-channel matrix, which can adapt the location to the global information and enhance the neural network's prediction capabilities. By spatially capturing the interaction between the current query location and all other locations in relation to the global information, the multi-channel matrix, referred to as location-aware mapped-representation and denoted as $R_{h,w}^{\mathrm{loc}}$ for each $x_{h,w}$, becomes an integral part of the overall framework.

However, due to the complex and high-dimensional nature of geo-spatial interactions, the distribution represented by the Location-aware mapped-representation is not easily discernible. Therefore, a simplified distribution is assumed as prior information for the location-aware mapped-representation. Specifically, in the initial state, the influence of the query location on the entire map is assumed to spread from the location to its surroundings in a two-dimensional Gaussian distribution. This means that the probability of each location being affected by the query location follows a two-dimensional Gaussian distribution. Consequently, for each query location $x_{h,w}$, the affected probability value of the grid at location (h', w') can be determined, which is denoted as $m'_{(h,w)\to(h',w')}$, and

$$m'_{(h,w)\to(h',w')} = P\,((a,b) = x_{h',w'}) \qquad (4)$$

where (a, b) are two-dimensional random variables and $(a, b) \sim \mathcal{N}(x_{h,w}, \Sigma)$ represents that (a, b) follows a normal distribution centered around the point $x_{h,w}$ on the map with covariance matrix Σ. As a result, for each query location $x_{h,w}$, a matrix of the same size as the original gridded map can be obtained,

$$M'_{h,w} = \begin{bmatrix} m'_{(h,w)\to(1,1)} & \cdots & m'_{(h,w)\to(1,W)} \\ \vdots & \ddots & \vdots \\ m'_{(h,w)\to(H,1)} & \cdots & m'_{(h,w)\to(h,w)} \end{bmatrix} \qquad (5)$$

Then $M'_{h,w}$ can be normalized for constructing a Gaussian location mask $M_{h,w}$. Based on the Gaussian location mask as a prior, the Location-aware Mapped-representation can be learned as a posterior via a convolutional neural network and be represented as a neural vector $f^{\mathrm{map}}(\cdot)$ finally, and the forward process is formulated as

$$R^{\mathrm{loc}}_{h,w} = f^{\mathrm{map}}(M_{h,w}),\ h \in [1, H],\ w \in [1, W] \tag{6}$$

3.2.5 Global-temporal-location representations

The framework consists of three main representations: the global-aware meta-representation (R^{meta}), the temporal-aware speed-representation ($R^{\mathrm{speed}}_{h,w}$), and the location-aware mapped-representation ($R^{\mathrm{loc}}_{h,w}$).

In this framework, each of these representations is assigned a different size as a neural vector representation. Additionally, all three representations have a one-dimensional channel. This means that for each grid in a query location, there are three one-dimensional vectors representing global reference data, temporal speed data, and spatial relations. This allows these vectors to carry information for making predictions.

The global-aware meta-representation specifically represents all global reference information. As a result, it can be used as a generalized feature map and has the potential to complete multiple tasks.

3.2.6 RFPM

In order to obtain accurate prediction results for a given query location $x_{h,w}$, a concatenation fusion technique is employed on three key representations: global-aware meta-representation (R^{meta}), temporal-aware speed-representation ($R^{\mathrm{speed}}_{h,w}$), and location-aware mapped-representation ($R^{\mathrm{loc}}_{h,w}$). The global-aware meta-representation (R^{meta}) captures high-level information about the overall context and global patterns in the data. It provides a broad understanding of the underlying factors that influence the prediction results. The temporal-aware speed-representation ($R^{\mathrm{speed}}_{h,w}$) focuses on the temporal dynamics and speed-related features of the data. It takes into account the changes and trends over time, which are crucial for accurate predictions. The location-aware mapped-representation ($R^{\mathrm{loc}}_{h,w}$) incorporates spatial information and maps the query location to its corresponding features. This representation enables the model to consider the specific characteristics and context of the query location.

By concatenating these three representations, the fusion process combines their unique strengths and enhances the overall predictive power. The concatenated representation is then fed into a GRU with a fully connected output layer, which are collectively referred to as $f^{\mathrm{GRU}}_{h,w}(\cdot)$. The GRU[58] is a type of recurrent neural network that effectively captures sequential dependencies in the data, and it processes the concatenated representation and learns to extract relevant patterns and relationships. Specifically, the input to the GRU network is a concatenation vector of three different spatio-temporal information,

$$c = \mathrm{concat}\,[R^{\mathrm{meta}}, R^{\mathrm{speed}}_{h,w}, R^{\mathrm{loc}}_{h,w}] \tag{7}$$

The GRU network then performs the task of predicting the temporal ordering of this fused concatenation vector.

Finally, the output of the GRU is passed through a fully connected output layer, since $y'_{h,w}$ has a different length with the final output that needs to be predicted. That applies appropriate transformations and computations to generate the prediction results $\hat{y}^{L'}_{h,w}$ for the query location $x_{h,w}$. This output is a multi-dimensional vector, where each element represents the speed value at a specific time point in the monitored time series. To clarify, the dimension of this vector corresponds to the length of the time series to be predicted, denoted as L'. There is no overlap time series between the $\hat{y}^{L'}_{h,w}$ and $V^{L}_{h,w}$.

The forward propagation process of representation fusion and prediction module can be described as follows:

$$\hat{y}^{L'}_{h,w} = f^{\mathrm{GRU}}_{h,w}(c) \tag{8}$$

Overall, the concatenation fusion technique, combined with the GRU and fully connected output layer, enables the model to effectively leverage the global-aware meta-representation, temporal-aware speed-representation, and location-aware mapped-representation to generate accurate and detailed prediction results for the given query location.

3.2.7 Optimization and inference

When the model is trained, each query location $x_{h,w}$ produces a temporal sequence of traffic speed values $\hat{y}^{L'}_{h,w}$, consisting of L' time points. This sequence is denoted as $\hat{y}^{L'}_{h,w} = [\hat{y}^{1}_{h,w}, \hat{y}^{2}_{h,w}, \ldots, \hat{y}^{l}_{h,w}]$. To measure the loss between $\hat{y}^{L'}_{h,w}$ and the ground truth $y^{L'}_{h,w}$, the Mean Squared Error (MSE) is used as the objective function. The ground truth sequence is represented as $y^{L'}_{h,w} = [y^{1}_{h,w}, y^{2}_{h,w}, \ldots, y^{l}_{h,w}]$. Thus, the loss can be formulated as follows:

$$\mathrm{Loss}^{L'}_{h,w} = \frac{1}{L'} \sum_{l}^{L'} \left\| \hat{y}^{L'}_{h,w} - y^{L'}_{h,w} \right\|_2^2 \tag{9}$$

The neural networks in MFGM, SEM, QLMM, and RFPM are trained to converge step by step using the aforementioned objective function. Due to the diverse and complex nature of global reference information, MFGM is designed to be bulky and heavy. However, once the model converges, the global-aware meta-representation is preserved and can effectively capture the extensive global reference information. Consequently, after the training process is completed, only the lightweight neural networks in SEM, QLMM and RFPM are utilized for fast and efficient inference. Specifically, the location-aware mapped-representation generated by QLMM and temporal-aware speed-representation achieved by SEM are directly fused with the saved global-aware meta-representation and fed into RFPM to accomplish the prediction task.

4 Experiment

4.1 Implementation details

The experiments in this study utilize datasets obtained from Refs. [59–61]. The datasets consist of a comprehensive collection of data related to a specific city, including 28 550 real estate listings, 497 256 points of interest, 250 000 traffic records, and over 100 million pieces of geolocated social media text. To effectively utilize the different types of data, namely real estate listings, social media text, and points of interest, they are considered as the global reference information. Prior to conducting the experiments, the reference information undergoes a preprocessing step to extract multiple metrics that can be used for analysis.

● **Social media text:** All the retrieved natural language texts from social media platforms are fed into a pre-trained Bidirectional Encoder Representations from Transformers (BERT)[62] model. This process generates a multi-dimensional vector for each grid cell, representing the social media text information. Subsequently, a pre-trained auto-encoder[63] is employed to reduce the dimensionality of the obtained vectors. The result is a multi-channel matrix that serves as part of the training data for the experiments.

● **Real estates:** The average price of all real estate listings within each grid cell is considered as an indicator of the living quality in that area. To represent this information, a normalized average real estate price is used as the real estate vector for each grid cell. In cases where there is no real estate information available for a particular grid cell, a value of zero is assigned. Ultimately, a single-channel matrix is obtained for the entire gridded map, representing the real estate features.

● **Points of interest:** This part of data consists of 23 different categories, which collectively provide diverse information about the interests and amenities present in each grid cell. To represent this information, a count-based one-hot encoding technique is applied to vectorize the points of interest within each grid cell. This results in a multi-channel matrix that represents the points of interest features for the entire gridded map.

By employing these preprocessing techniques and representing the reference information in various formats, the experiments aim to leverage the rich and diverse data available to gain insights and make informed analyses.

The traffic speed values used in the model are categorized into two parts: inputs and outputs. To provide a comprehensive overview, the entire map consists of approximately 250 000 grids, out of which around 24 000 grids have recorded traffic data. Each grid contains traffic data for 216 time points, covering from 6:00 a.m. to 12:00 a.m. In order to establish input-output pairs for the traffic speed values, 27 time points (equivalent to 45 minutes) are selected as the input, while 9 time points (equivalent to 15 minutes) are designated as the output. These pairs are utilized for training, validation, and testing purposes. The division of the traffic speed values into these sets is based on geographic locations, and the results obtained from the testing set are reported. To ensure an appropriate distribution, all grids are divided into three sets: training set, validation set, and test set. The ratio of this division is 60% for the training set, 20% for the validation set, and 20% for the test set. This dataset split is commonly accepted in machine learning and strikes a balance between having a sufficiently large test set for meaningful evaluation and ensuring a reasonable amount of data for training.

Furthermore, we conduct our experiment with a set of standard hyperparameters. The hyperparameters used are listed in Table 1. During the training process, the model tends to converge approximately after 10 epochs as shown in Fig. 3. As the problem at hand transitions from a classification task to a regression task, several metrics are employed to evaluate the

Table 1 Hyperparameter table.

Hyperparameter	Value
Batch size	16
Learning rate	1×10^{-3}
Learning rate decay	0.1
Training epochs	20
Loss function	MSE
Input sequence length	27
Ouput sequence length	9

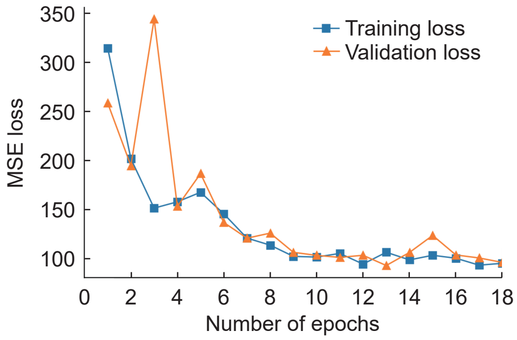


Fig. 3 Training loss curve and validation loss curve during model training.

models. These metrics include Mean Absolute Error (MAE), MSE, Root Mean Square Error (RMAE), and R-Squared (R^2). R^2, known as the coefficient of determination, is a statistical measure that represents the proportion of the variance in the dependent variable (target) that is predictable from the independent variables (features) in a regression model. An R^2 value closed to 1 indicates that the model’s predictions are highly correlated with the actual data, reflecting a model that captures the underlying traffic patterns effectively. Mathematically, R^2 is calculated using the following formula:

$$R^2 = 1 - \frac{\text{Sum of squared residuals}}{\text{Total sum of squares}} \tag{10}$$

4.2 Results

After training and validating the entire framework on the selected dataset, quantitative results are obtained from the evaluation on the testing set, as shown in Table 2. Since our framework considers the relationship between various spatial modalities and temporal speed information, which is currently with limited innovation in the field of traffic prediction, we have conducted comparative experiments from separate spatial and temporal perspectives. For the temporal aspect, RNN-based traffic prediction method is employed, with RNN, GRU[58], and LSTM[64] serving as three kinds of baseline. For each one, we conduct experiments with hidden layers set at 8, 16, and 32. Specifically, given speed values of a fixed time period and a query location, these baselines could output subsequent speed values on that geo-location. For the spatial aspect, MFRM[44] is used as the baseline. In detail, the MFRM network only utilizes global-aware meta-representation and location-aware mapped-representation, excluding the temporal-aware speed-representation compared to our approach. As indicated in Table 2, all metrics of the main experiment outperform the baselines, thus demonstrating the effectiveness of the proposed framework. Our framework boasts the highest R^2 score in comparison to other baseline models, indicating that it effectively captures and explains all variations in the target variable, showcasing its robust predictive capabilities.

Table 2 Evaluation results on our method and baseline methods. MAE, MSE, and RMSE are arranged in descending order, where lower values indicate better model performance; while R^2 is arranged in ascending order, where higher values indicate better model performance.

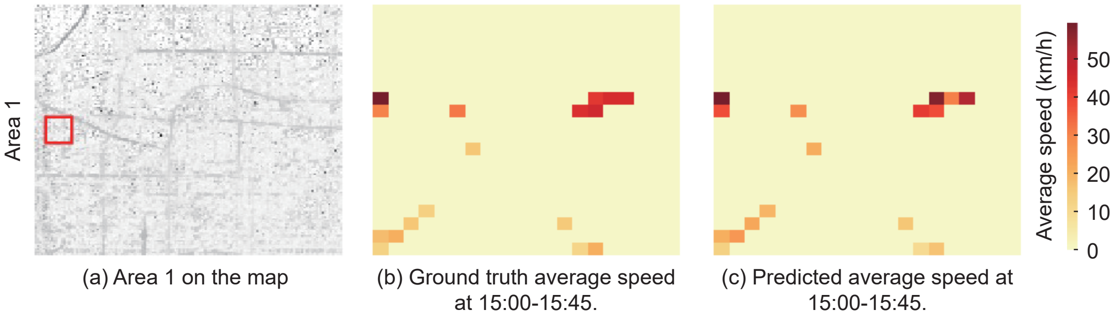
Model	MAE ↓	MSE ↓	RMSE ↓	R^2 ↑
MFRM[44]	13.4479	281.5114	16.7783	0.1828
RNN (*h*=8)	9.9929	175.7345	13.2565	0.5009
RNN (*h*=16)	9.4586	158.9140	12.6061	0.5487
RNN (*h*=32)	9.3465	153.7023	12.3977	0.5635
LSTM (*h*=8)	9.5288	161.7115	12.7166	0.5407
LSTM (*h*=16)	8.4397	125.9219	11.2215	0.6424
LSTM (*h*=32)	8.3978	123.5179	11.1139	0.6492
GRU (*h*=8)	8.6617	135.2399	11.6293	0.6159
GRU (*h*=16)	8.4636	129.7345	11.3901	0.6316
GRU (*h*=32)	8.1421	117.7126	10.8495	0.6657
Ours	**7.1297**	**91.4398**	**9.5624**	**0.7403**

Furthermore, it can be observed that the addition of temporal information improves the performance of the network compared to the MFRM[44], which solely relies on spatial information from the map for traffic speed prediction. This suggests that our model successfully integrates both spatial and temporal information, adding an extra dimension to the original modality. The comparison between the baseline results and the main experimental results highlights the difficulty in obtaining meaningful prediction outcomes by solely relying on the simple geographical relationship between different roads on a large-scale map. This further validates the necessity and effectiveness of the proposed framework.

In addition, qualitative results are also presented through Figs. 4 and 5. These images are selected to represent two different areas on the map with two different time periods. Figure 4 displays the ground truth and predicted average speed values from 15:00 p.m. to 15:45 p.m. Figure 5 shows the ground truth and predicted average speed values from 19:30 p.m. to 20:15 p.m. The selection of these two time points allows for the observation of different traffic flows during less busy hours in the afternoon and evening rush hours. It is important to note that the specific speed values shown in the images are the average values of traffic speed flow within the given time range at each location. By comparing the observed real speed values with the predicted values during the same time period, it can be observed that the ranges of values are largely consistent in many locations. Despite the differences in color representation, the differences in value range are generally within 10 km/h. This demonstrates the robustness and accuracy of our network.

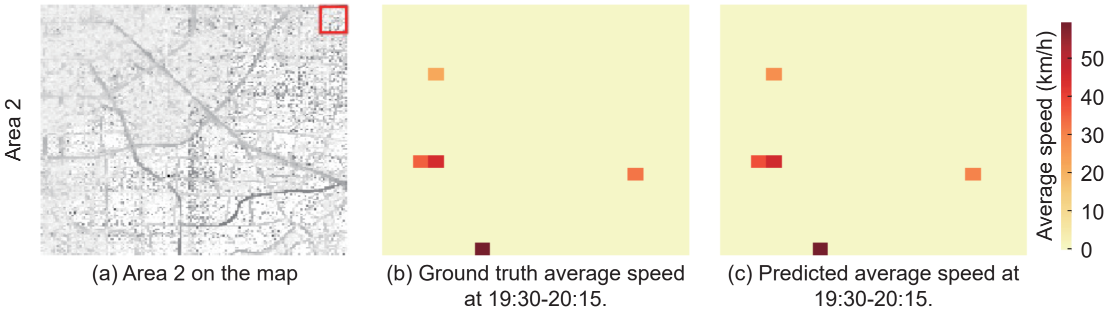
4.3 Ablation study

In the eventual fusion stage (i.e., RFPM) of the framework, the global-aware meta-representation is deactivated by replacing the original fully connected layer with a zero-matrix. Similarly, when the location-aware mapped-representation and temporal-aware speed-representation are deactivated, their respective zero-matrices are also incorporated into the network, which effectively nullify their influence on the model's output. By deactivating these representations, the model focuses solely on other relevant features, allowed for a more targeted analysis of the remaining representations and enhances the model's ability to capture and interpret specific patterns and relationships within the data.



(a) Area 1 on the map (b) Ground truth average speed at 15:00-15:45. (c) Predicted average speed at 15:00-15:45.

Fig. 4 Qualitative results, (a) shows Area 1, (b) shows ground truth average speed values from 15:00 to 15:45, and (c) shows predicted average speed values from 15:00 to 15:45



(a) Area 2 on the map (b) Ground truth average speed at 19:30-20:15. (c) Predicted average speed at 19:30-20:15.

Fig. 5 Qualitative results, (a) shows Area 2, (b) shows ground truth average speed values from 19:30 to 20:15, and (c) shows predicted average speed values from 19:30 to 20:15

The evaluation results, as shown in Table 3, indicate that not only do the three modules of the model perform individually, but also their combined performance falls behind the network structure where all three modules are integrated together. This clearly demonstrates the indispensability of the three parts of representations in the proposed framework. It is noteworthy that the posterior information learned from the SEM, plays a crucial role in the performance. Additionally, the reference information included in global-aware meta-representation and location-aware mapped-representation also proves to be decisive in making accurate predictions.

4.4 Inference performance analysis

We have conducted a comparative analysis of our model's inference speed and performance, as illustrated in Fig. 6. Specifically, the vertical axis in the graph represents the model's MAE accuracy, while the horizontal axis signifies the model's inference time for each inference. Our model exhibits superior performance, with its inference time being only marginally slower than that of LSTM or GRU networks by less than 20 ms. Although the MFRM network shares a similar structure with our model, including features of global-aware meta-representation and location-aware mapped-representation, its inference time surpasses 100 ms. This clearly demonstrates that our model effectively balances inference speed and performance.

Moreover, we have analyzed the complexity associated with our proposed method. In our experiments, varying proportions of data are extracted from the test dataset to serve as inputs for the model (20%, 40%, 60%, 80%, and 100%). As depicted in Fig. 7, a obvious linear correlation emerges between the time spent on model inference and the volume of input data, establishing our algorithm's time complexity as $O(n)$. Simultaneously, we observe the GPU memory consumption during inference experiments, revealing a fixed value during inference. This underscores our algorithm's space complexity, which stands at $O(1)$.

5 Conclusion

We present an innovative method for large-scale traffic speed prediction by integrating both spatial and temporal information. By effectively representing the

Table 3 Evaluation results with different representations. MAE, MSE, and RMSE are arranged in descending order, where lower values indicate better model performance; while R^2 is arranged in ascending order, where higher values indicate better model performance.

Global-aware meta-representation	Location-aware mapped-representation	Temporal-aware speed-representation	MAE ↓	MSE ↓	RMSE ↓	R^2 ↑
✓	–	–	22.4678	675.2606	25.9858	−0.9181
–	✓	–	21.2648	608.7488	24.6728	−0.7292
–	–	✓	7.1700	92.3510	9.6099	0.7377
✓	✓	–	22.2869	665.1166	25.7899	−0.8893
✓	–	✓	7.1379	91.7397	9.5781	0.7395
–	✓	✓	22.2823	664.9396	25.7864	−0.8888
✓	✓	✓	**7.1201**	**91.2136**	**9.5506**	**0.7410**

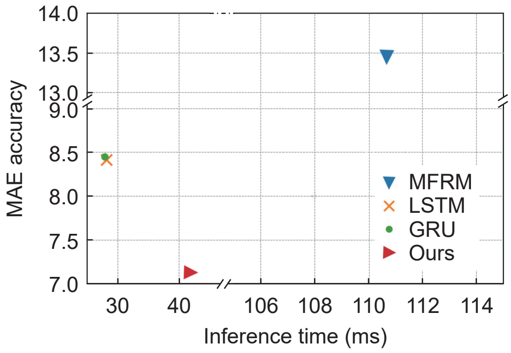


Fig. 6 Inference time and performance comparison among our method and baseline methods.

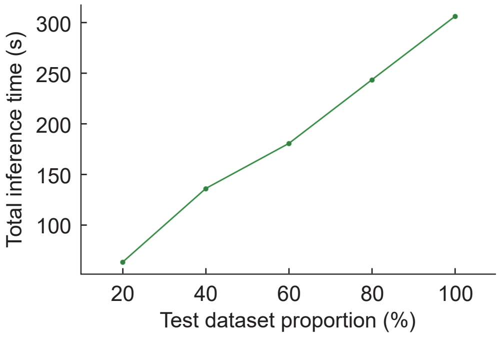


Fig. 7 Time complexity analysis based on different proportions of test dataset as input.

intricate interplay between spatial and temporal features within traffic data, the proposed method offers a comprehensive integration of abundant information available on extensive maps and enable precise inference on traffic conditions. Moreover, we implement our method through extensive experiments using real-world large-scale map data, demonstrating its superior performance compared to existing methods.

There are numerous directions for our work to explore in the future. Firstly, our current methodology is based on learning from several global reference information. However, in our implementation, we do not consider mixing different types of global reference information in different proportions. If this can be experimented with in future work, better results can be obtained. Secondly, our method is focusing on global map features. So we may consider narrowing the receptive field to localized scopes, which allows network to understand and interpret data better based on the features learned. Thirdly, the dataset contains multiple modalities. It would be worthwhile to consider aligning these modalities with each other. This approach could facilitate the development of a comprehensive model to predict across different modalities.

In summary, this paper contributes to the field of traffic prediction by introducing a multimodal fusion framework. The proposed approach and the network model hold substantial potential in real-world traffic planning and management.

References

[1] China Association of Automobile Manufacturers, Motor vehicle ownership reaches 430 million in China, http://www.caam.org.cn/chn/7/cate_120/con_5236191.html, 2023.

[2] K. Nellore and G. P. Hancke, A survey on urban traffic management system using wireless sensor networks, *Sensors*, vol. 16, no. 2, p. 157, 2016.

[3] J. Rios-Torres and A. A. Malikopoulos, Automated and cooperative vehicle merging at highway on-ramps, *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 4, pp. 780–789, 2017.

[4] Q. Shi and M. Abdel-Aty, Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways, *Transp. Res. Part C Emerg. Technol.*, vol. 58, pp. 380–394, 2015.

[5] S. Muthuramalingam, A. Bharathi, S. Rakesh Kumar, N. Gayathri, R. Sathiyaraj, and B. Balamurugan, IoT based intelligent transportation system (IoT-ITS) for global perspective: A case study, in *Internet of Things and Big Data Analytics for Smart Generation*, V. E. Balas, V. K. Solanki, R. Kumar, and M. Khari, eds. Cham, Germany: Springer, 2019, pp. 279–300.

[6] Z. Lv and W. Shang, Impacts of intelligent transportation systems on energy conservation and emission reduction of transport systems: A comprehensive review, *Green Technol. Sustain.*, vol. 1, no. 1, p. 100002, 2023.

[7] J. Liu, C. Chang, J. Liu, X. Wu, L. Ma, and X. Qi, MarS3D: A plug-and-play motion-aware model for semantic segmentation on multi-scan 3D point clouds, in *Proc. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Vancouver, Canada, 2023, pp. 9372–9381.

[8] Y. Zhao, H. Zheng, Z. Wang, J. Luo, and E. Y. Lam, Point cloud denoising via momentum ascent in gradient fields, in *Proc. 2023 IEEE Int. Conf. Image Processing*, Kuala Lumpur, Malaysia, 2023, pp. 161–165.

[9] X. Lai, Y. Chen, F. Lu, J. Liu, and J. Jia, Spherical transformer for LiDAR-based 3D recognition, in *Proc. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Vancouver, Canada, 2023, pp. 17545–17555.

[10] J. Liu, Y. Chen, X. Ye, Z Tian, X. Tan, and X. Qi, Spatial pruned sparse convolution for efficient 3D object detection, in *Proc. 36th Conf. Neural Information Processing Systems*, New Orleans, LA, USA, 2022, pp. 6735–6748.

[11] J. Yang, S. Shi, Z. Wang, H. Li, and X. Qi, ST3D: Self-training for unsupervised domain adaptation on 3D object detection, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 10363–10373.

[12] A. Faramarzi, M. Heidarinejad, B. Stephens, and S. Mirjalili, Equilibrium optimizer: A novel optimization algorithm, *Knowledge-Based Systems*, vol. 191, p. 105190, 2020.

[13] M. C. Popescu, V. Balas, L. P. Popescu, and N. Mastorakis, Multilayer perceptron and neural networks, *WSEAS Transactions on Circuits and Systems*, vol. 8, no.7, pp. 579–588, 2009.

[14] B. M. Williams and L. A. Hoel, Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results, *J. Transp. Eng.*, vol. 129, no. 6, pp. 664–672, 2003.

[15] Y. Zhang and Y. C. Liu, Traffic forecasting using least squares support vector machines, *Transportmetrica*, vol. 5, no. 3, pp. 193–213, 2009.

[16] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, Long short-term memory neural network for traffic speed prediction using remote microwave sensor data, *Transp. Res. Part C Emerg. Technol.*, vol. 54, pp. 187–197, 2015.

[17] J. Tang, F. Liu, Y. Zou, W. Zhang, and Y. Wang, An improved fuzzy neural network for traffic speed prediction considering periodic characteristic, *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 9, pp. 2340–2350, 2017.

[18] D. W. Xu, Y. D. Wang, L. M. Jia, Y. Qin, and H. H. Dong, Real-time road traffic state prediction based on ARIMA and Kalman filter, *Front. Inf. Technol. Electron.*

Eng., vol. 18, no. 2, pp. 287–302, 2017.

[19] H. Yang, L. Du, G. Zhang, and T. Ma, A traffic flow dependency and dynamics based deep learning aided approach for network-wide traffic speed propagation prediction, *Transp. Res. Part B Methodol.*, vol. 167, pp. 99–117, 2023.

[20] M. Levin and Y. D. Tsao, On forecasting freeway occupancies and volumes (abridgment), *Transp. Res. Record*, vol. 773, pp. 47–49, 1980.

[21] T. Bogaerts, A. D. Masegosa, J. S. Angarita-Zapata, E. Onieva, and P. Hellinckx, A graph CNN-LSTM neural network for short and long-term traffic forecasting based on trajectory data, *Transp. Res. Part C Emerg. Technol.*, vol. 112, pp. 62–77, 2020.

[22] M. S. Ahmed and A. R. Cook, Analysis of freeway traffic time-series data by using Box-Jenkins techniques, *Transp. Res. Record*, vol. 722, pp. 1–9, 1979.

[23] D. A. Tedjopurnomo, Z. Bao, B. Zheng, F. M. Choudhury, and A. K. Qin, A survey on modern deep neural network for traffic prediction: Trends, methods and challenges, *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 4, pp. 1544–1561, 2022.

[24] Y. Zhao, G. Li, and E. Y. Lam, Cross-camera human motion transfer by time series analysis, in *Proc. 2024 IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Seoul, Republic of Korea, 2024, pp. 4985–4989.

[25] C. De Fabritiis, R. Ragona, and G. Valenti, Traffic estimation and prediction based on real time floating car data, in *Proc. 2008 11th Int. IEEE Conf. Intelligent Transportation Systems*, Beijing, China, 2008, pp. 197–203.

[26] Z. Liu, Z. Li, K. Wu, and M. Li, Urban traffic prediction from mobility data using deep learning, *IEEE Network*, vol. 32, no. 4, pp. 40–46, 2018.

[27] P. P. Dubey and P. Borkar, Review on techniques for traffic jam detection and congestion avoidance, in *Proc. 2015 2nd Int. Conf. Electronics and Communication Systems*, Coimbatore, India, 2015, pp. 434–440.

[28] Y. Yan, Y. Deng, S. Cui, Y. H. Kuo, A. H. F. Chow, and C. Ying, A policy gradient approach to solving dynamic assignment problem for on-site service delivery, *Trans. Res. Part E Logist. Transp. Rev.*, vol. 178, p. 103260, 2023.

[29] Y. Yan, A. H. F. Chow, C. P. Ho, Y. H. Kuo, Q. Wu, and C. Ying, Reinforcement learning for logistics and supply chain management: Methodologies, state of the art, and future opportunities, *Transp. Res. Part E Logist. Transp. Rev.*, vol. 162, p. 102712, 2022.

[30] Z. Ning, H. Chen, E. C. H. Ngai, X. Wang, L. Guo, and J. Liu, Lightweight imitation learning for real-time cooperative service migration, *IEEE Trans. Mobile Comput.*, vol. 23, no. 2, pp. 1503–1520, 2024.

[31] H. Chen, X. Wang, Z. Ning, and L. Guo, SDN-enabled 3C resource integration in green internet of electrical vehicles, in *Proc. of CECNet 2021*, doi:10.3233/FAIA210443.

[32] Y. Zhao, Q. Zeng, and E. Y. Lam, Adaptive compressed sensing for real-time video compression, transmission, and reconstruction, in *Proc. 2023 IEEE 10th Int. Conf. Data Science and Advanced Analytics*, Thessaloniki, Greece, 2023, pp. 1–10.

[33] Y. Zhao and E. Y. Lam, SASA: Saliency-aware self-adaptive snapshot compressive imaging, in *Proc. 2024 IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Seoul, Republic of Korea, 2024, pp. 2370–2374.

[34] Y. Zhao, H. Zheng, J. Luo, and E. Y. Lam, Improving video colorization by test-time tuning, in *Proc. 2023 IEEE Int. Conf. Image Processing*, Kuala Lumpur, Malaysia, 2023, pp. 166–170.

[35] Y. Qi and S. Ishak, A hidden Markov model for short term prediction of traffic conditions on freeways, *Transp. Res. Part C Emerg. Technol.*, vol. 43, pp. 95–111, 2014.

[36] B. Yu, H. Yin, and Z. Zhu, Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. in *Proc. 27th Int. Joint Conf. Artificial Intelligence*, Stockholm, Sweden, 2018, pp. 3634–3640.

[37] R. Vinayakumar, K. P. Soman, and P. Poornachandran, Applying deep learning approaches for network traffic prediction, in *Proc. 2017 Int. Conf. Advances in Computing, Communications and Informatics*, Udupi, India, 2017, pp. 2353–2358.

[38] Y. Tian and L. Pan, Predicting short-term traffic flow by long short-term memory recurrent neural network, in *Proc. 2015 IEEE Int. Conf. Smart City/SocialCom/SustainCom*, Chengdu, China, 2015, pp. 153–158.

[39] Z. Pan, Y. Liang, W. Wang, Y. Yu, Y. Zheng, and J. Zhang, Urban traffic prediction from spatio-temporal data using deep meta learning, in *Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining*, Anchorage, AK, USA, 2019, pp. 1720–1730.

[40] Y. Zhao, H. Zheng, Z. Wang, J. Luo, and E. Y. Lam, MANet: Improving video denoising with a multi-alignment network, in *Proc. 2022 IEEE Int. Conf. Image Processing*, Bordeaux, France, 2022, pp. 2036–2040.

[41] S. Zhao, R. Y. Zhong, J. Wang, C. Xu, and J. Zhang, Unsupervised fabric defects detection based on spatial domain saliency and features clustering, *Comput. Ind. Eng.*, vol. 185, p. 109681, 2023.

[42] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, Attention based spatial-temporal graph convolutional networks for traffic flow forecasting, in *Proc. 33rd AAAI Conf. Artificial Intelligence*, Honolulu, HI, USA, 2019, pp. 922–929.

[43] Y. Zhao, Z. Wang, and E. Y. Lam, Improving source localization by perturbing graph diffusion, in *Proc. 2022 IEEE 9th Int. Conf. Data Science and Advanced Analytics*, Shenzhen, China, 2022, pp. 1–9.

[44] B. Zhou, J. Liu, S. Cui, and Y. Zhao, Large-scale traffic congestion prediction based on multimodal fusion and representation mapping, in *Proc. 2022 IEEE 9th Int. Conf. Data Science and Advanced Analytics*, Shenzhen, China, 2022, pp. 1–9.

[45] Z. Zhou, X. Dong, Z. Li, K. Yu, C. Ding, and Y. Yang, Spatio-temporal feature encoding for traffic accident

detection in VANET environment, *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 19772–19781, 2022.

[46] Y. Li, R. Yu, C. Shahabi, and Y. Liu, Diffusion convolutional recurrent neural network: Data-driven traffic forecasting, in *Proc. of 6th Int. Conf. Learning Representations*, Vancouver, Canada, doi:10.48550/arXiv. 1707.01926.

[47] X. Wang, C. Chen, Y. Min, J. He, B. Yang, and Y. Zhang, Efficient metropolitan traffic prediction based on graph recurrent neural network, arXiv preprint arXiv: 1811.00740, 2018.

[48] K. Zhang, L. Zheng, Z. Liu, and N. Jia, A deep learning based multitask model for network-wide traffic speed prediction, *Neurocomputing*, vol. 396, pp. 438–450, 2020.

[49] T. Nguyen, G. Nguyen, and B. M. Nguyen, EO-CNN: An enhanced CNN model trained by equilibrium optimization for traffic transportation prediction, *Procedia Comput. Sci*., vol. 176, pp. 800–809, 2020.

[50] D. Yang, S. Li, Z. Peng, P. Wang, J. Wang, and H. Yang, MF-CNN: Traffic flow prediction using convolutional neural network and multi-features fusion, *IEICE Trans. Inf. Syst*., vol. E102, no. 8, pp. 1526–1536, 2019.

[51] M. Z. Mehdi, H. M. Kammoun, N. G. Benayed, D. Sellami, and A. D. Masmoudi, Entropy-based traffic flow labeling for CNN-based traffic congestion prediction from meta-parameters, *IEEE Access*, vol. 10, pp. 16123–16133, 2022.

[52] J. Wang, S. Zhao, C. Xu, J. Zhang, and R. Zhong, Brain-inspired interpretable network pruning for smart vision-based defect detection equipment, *IEEE Trans. Ind. Inf.*, vol. 19, no. 2, pp. 1666–1673, 2023.

[53] W. Min and L. Wynter, Real-time road traffic prediction with spatio-temporal correlations, *Transp. Res. Part C Emerg. Technol.*, vol. 19, no. 4, pp. 606–616, 2011.

[54] H. Yu, Z. Wu, S. Wang, Y. Wang, and X. Ma, Spatiotemporal recurrent convolutional networks for traffic prediction in transportation networks, *Sensors*, vol. 17, no. 7, p. 1501, 2017.

[55] C. Qiu, Y. Zhang, Z. Feng, P. Zhang, and S. Cui, Spatio-temporal wireless traffic prediction with recurrent neural network, *IEEE Wirel. Commun. Lett.*, vol. 7, no. 4, pp. 554–557, 2018.

[56] X. Luo, D. Li, Y. Yang, and S. Zhang, Spatiotemporal traffic flow prediction with KNN and LSTM, *J. Adv. Transp.*, vol. 2019, p. 4145353, 2019.

[57] J. Wang, R. Chen, and Z. He, Traffic speed prediction for urban transportation network: A path based deep learning approach, *Transportation Research Part C*: *Emerging Technologies*, vol. 100, pp. 372–385, 2019.

[58] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, arXiv preprint arXiv: 1412.3555, 2014.

[59] Y. Zhao, S. Shi, R. Ravi, Z. Wang, E. Y. Lam, and J. Zhao, H4M: Heterogeneous, multi-source, multi-modal, multi-view and multi-distributional dataset for socioeconomic analytics in the case of Beijing, in *Proc. 2022 IEEE 9th Int. Conf. Data Science and Advanced Analytics*, Shenzhen, China, 2022, pp. 1–10.

[60] Y. Zhao, R. Ravi, S. Shi, Z. Wang, E. Y. Lam, and J. Zhao, PATE: Property, amenities, traffic and emotions coming together for real estate price prediction, in *Proc. 2022 IEEE 9th Int. Conf. Data Science and Advanced Analytics*, Shenzhen, China, 2022, pp. 1–10.

[61] Y. Zhao, J. Zhao, and E. Y. Lam, House price prediction: A multi-source data fusion perspective, *Big Data Mining and Analytics*, doi:10.26599/BDMA.2024.9020019.

[62] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies, Volume 1*, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[63] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Learning internal representations by error propagation, in *Parallel Distributed Processing*: *Explorations in the Microstructures of Cognition*, D. E. Rumelhart and J. L. McClelland, eds. Cambridge, MA, USA: MIT Press, 1986, pp. 318–362.

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

Bodong Zhou received the BS degree from East China University of Science and Technology, China in 2019, and the MEng degree from Northeastern University, USA in 2021. He is currently serving as the independent technical expert at Technical Consulting Department, Shanghai EchoBlend Internet Technology Co. Ltd., China. His current research interests lie in large language models, multimodal learning, and diffusion models, especially in video generation.

Jiahui Liu received the BS degree from East China University of Science and Technology, China in 2019, and the MEng degree from University of Sheffield, UK in 2020. He is currently a PhD candidate at the University of Hong Kong, China. His research interests include representation learning and computer vision.

Yaping Zhao received the BS degree from Beihang University, China in 2018, and the MS degree from Tsinghua University, China in 2021. She is currently a PhD candidate at Department of Electrical and Electronic Engineering, the University of Hong Kong (HKU), China. Her research interests encompass computational imaging, computer vision, machine learning, and artificial intelligence. She has been recognized with several awards, including the Research Postgraduate Student Innovation Award (HKU), AAAI-23 student scholarship, IEEE SPS Travel Grant, IEEE CIS Conference Participation Grant, s-EDSSC Best Student Paper Award, and the WISE Best Poster Award. She has also contributed as a research track chair for IEEE DSAA 2022, a special session organizer for IEEE DSAA since 2022, a program committee member of AAAI and CICAI, and a reviewer of *CVPR*, *ECCV*, *WACV*, *ACCV*, *IEEE TIP*, and *TNNLS*.

Songyi Cui received the BEng degree in traffic engineering from Beijing University of Technology, China in 2020, and the MEng degree in industrial engineering and logistics management from the University of Hong Kong, China in 2021. She is currently a PhD candidate at Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong, China. Her research interests include traffic optimization, traffic simulation, and deep learning applications in transportation systems.