# ESTNet: Embedded Spatial-Temporal Network for Modeling Traffic Flow Dynamics

Guiyang Luo, *Graduate Student Member, IEEE*, Hui Zhang, Quan Yuan, *Member, IEEE*, Jinglin Li, *Member, IEEE*, and Fei-Yue Wang, *Fellow, IEEE*

***Abstract*— Accurate spatial-temporal prediction is a fundamental building block of many real-world applications such as traffic scheduling and management, environment policy making, and public safety. This problem is still challenging due to nonlinear, complicated, and dynamic spatial-temporal dependencies. To address these challenges, we propose a novel embedded spatial-temporal network (ESTNet), which extracts efficient features to model the dynamic correlations and then exploits three-dimension convolution to synchronously model the spatial-temporal dependencies. Specifically, we propose multi-range graph convolution networks for extracting multi-scale static features from the fine-grained road network. Meanwhile, dynamic features are extracted from real-time traffic using a gated recurrent unit network. These features can be applied to identify the dynamic and flexible correlations among sensors and make it possible to exploit a three-dimension convolution unit (3DCon) to simultaneously model the spatial-temporal dependencies. Furthermore, we propose a residual network by stacking multiple 3DCon to capture the nonlinear and complicated dependencies. The effectiveness and superiority of ESTNet are verified on two real-world datasets, and experiments show ESTNet outperforms the state-of-the-art with a significant margin. The code and models will be publicly available.**

***Index Terms*— Traffic forecasting, graph convolutional network, spatial-temporal networks.**

## I. Introduction

RECENT years have witnessed a sharp rise in the popularity and prosperity of spatial-temporal prediction in intelligent transportation systems, due to the growing amount of traffic-related datasets and its indispensable role in real-world applications [1], [2]. An accurate spatial-temporal prediction model is of great importance to a tremendous number of applications ranging from urban planning, traffic control to public safety [3]. For example, it can assist the traffic management bureau in making proactive and dynamic traffic control decisions, help drivers with intelligent route guidance, and alleviate the huge congestion problem [4].

Great efforts have been devoted to improving the spatial-temporal prediction accuracy, which is challenged by the complex, non-linear, and dynamic spatial-temporal dependencies [5]. Traditionally, basic time series models (e.g., ARIMA, Kalman Filtering and their variants), and regression models with spatial-temporal regularizations, are used for spatial-temporal prediction [6]. Recently, considerable attentions have been attracted by machine learning approaches, especially deep learning models [7]–[10], e.g., convolutional LSTM [11], ST-ResNet [12], etc. However, these models ignore the natural topology structure of the road network. For example, most existing works, e.g., DCRNN [5], Graph WaveNet [13], GMAN [4], apply the Euclidean distance between sensor locations to construct the adjacency matrix. However, the distance ignore the road topology between nodes, which may not reflect the true dependency between sensors. Based on the constructed adjacency matrix, graph convolution network (GCN) [14], which has the ability to handle irregular structured data, has been widely applied for the task of spatial-temporal prediction, e.g., DCRNN [5], Graph WaveNet [13], GMAN [4], SLC [15], AGCRN [16], etc.

Nevertheless, existing approaches are still inefficient and inaccurate in practice due to the following challenges:

1) **Dynamic spatial dependencies**. The correlations between locations are not static. For example, as shown in Fig.1(a), people commute to and from the workplace in the rush hours, resulting in heavy dynamic dependencies between the residential area and the business area. Social events (e.g., local museums and concerts, sporting events, outdoor festivals, city-wide tours) could lead to a sudden emergence of traffic flow at a place within a short period. Many existing approaches rely on pre-defined graphs using distance or similarity function [16] to learn the spatial dependencies between locations. However, this pre-defined graph is static and not applicable to capturing the spatial dynamics.

Manuscript received 21 August 2021; revised 29 November 2021 and 20 January 2022; accepted 14 March 2022. Date of publication 13 May 2022; date of current version 11 October 2022. This work was supported by the National Natural Science Foundation of China under Grant 62102041, Grant 61876023, and Grant 61902035. The Associate Editor for this article was L. Li. *(Corresponding author: Jinglin Li.)*

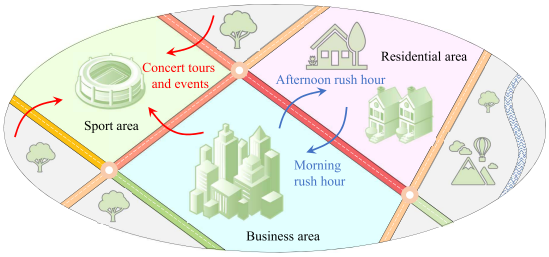
Guiyang Luo and Quan Yuan are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100088, China (e-mail: luoguiyang@bupt.edu.cn; yuanquan@bupt.edu.cn).

Hui Zhang is with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China, and also with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: huizhang1@bjtu.edu.cn).

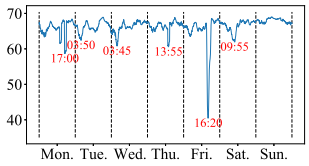
Jinglin Li is with the Science and Technology on Communication Networks Laboratory, Beijing University of Posts and Telecommunications, Beijing 100088, China, and also with the State Key Laboratory of Networking and Switching Technology, Shijiahzuang, Hebei 050081, China (e-mail: jlli@bupt.edu.cn).

Fei-Yue Wang is with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, also with the Innovation Center for Parallel Vision, Qingdao Academy of Intelligent Industries, Qingdao 266000, China, and also with the Institute of Systems Engineering, Macau University of Science and Technology, Macau, China (e-mail: feiyue@ieee.org).

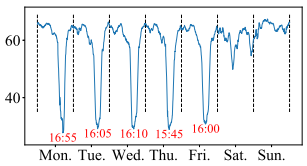
Digital Object Identifier 10.1109/TITS.2022.3167019




(a) Non-static spatial dependencies.




(b) Dynamic temporal correlations.




(c) Periodicity and temporal shifting.

Fig. 1. Dynamic correlations. (a) shows the non-static spatial dependencies caused by rush hours, and social events; (b) and (c) are the speed within one week of two different locations. (b) shows an abnormal dynamic temporal correlation; (c) shows the dynamic periodicity and temporal shifting.

2) **Dynamic temporal dependencies**. The temporal correlations are also dynamic. The social events could lead to an abnormally low speed, since a tremendous amount of cars are accumulated within a short period. As shown in Fig.1(b), sudden congestion is caused on Friday, leading to a change of temporal correlations. Besides, traffic data show a strong daily and weekly periodicity (temporal closeness, period, and trend properties [17]), as shown in Fig.1(c). However, these different kinds of periodicity show distinguishable temporal correlations. For example, the patterns between weekdays and weekends are different. Going a step further with the speed in seven days, temporal shifting exists between different days, as shown in Fig.1(c).
3) **Separate modeling of spatial and temporal correlation.** Many existing approaches exploit GCN to capture spatial dependencies while RNN or CNN for learning temporal correlations. Nevertheless, there is still a lack of effective methods to capture spatial and temporal dependencies synchronously. STSGCN [18] achieves it by connects individual spatial graphs of adjacent time steps into one graph. However, it suffers from long-sequence temporal dependencies due to the concatenated adjacency matrix.

To tackle these challenges, we propose an embedded spatial-temporal network (ESTNet) which could not only capture both the changing spatial dependencies and dynamic temporal correlations, but also simultaneously model the spatial-temporal correlations using a complete 3D convolutional network. The main contributions of this paper are as follows:

1) Instead of constructing the sensor graph using Euclidean similarity between sensors, we propose to extract it from the road network with fine-grained road topology and road segment attributes. Specifically, we exploit GCN networks with multiple ranges to extract multi-scale static features from a fine-grained road network based on vanilla GCN network. Combining the dynamic features extracted by gated recurrent unit (GRU) from real-time traffic data, it is possible to dynamically learn both the spatial and temporal correlations and then construct the sensor graph exploiting the similarity between the learned features.
2) A sequence encoder consisting of a GRU network is adopted to extract dynamic features for capturing dynamic correlations. Combining both static and dynamic features, we are able to dynamically select a fixed number of most correlated neighbors, making it possible to exploit a 3D convolutional unit to model the spatial-temporal correlations synchronously. Furthermore, we design a residual network to capture complex and nonlinear spatial-temporal correlations.
3) We evaluate our ESTNet on two real-world traffic datasets and the experimental results show that ESTNet outperforms the state-of-the-art with a significant margin, achieving an average 8.15% and 7.01% relative improvement to the existing best results on METR-LA and PEMS-BAY dataset, respectively.

The remainder of this paper is organized as follows: Section II reviews related works. Section III presents a formulation to spatial-temporal prediction. In Section IV, we introduce the ESTNet. The experimental evaluation and results are discussed in Section V. Finally, Section VI concludes this work.

## II. Related Works

Existing traffic forecasting approaches can be categorized as knowledge-driven and data-driven methods [19], [20], and more details can be found in the recent survey [3]. Recently, deep learning, and GCN based methods have dominated traffic prediction. The biggest difference between them lies in the regular or non-regular traffic data, i.e., deep learning methods [21]–[23] require the input be regular, and GCN methods could process the irregular traffic data directly.

### A. Deep Learning Based Prediction

Kang *et al.* [24] employ the long short-term memory (LSTM) recurrent neural network to analyze the effects of various input settings on the LSTM prediction performances. Liu *et al.* [25] propose Conv-LSTM to extract the spatial-temporal information of the traffic flow information, which combine convolution and LSTM. Zheng *et al.* [26] propose a deep and embedding learning approach, which take fine-grained traffic information, route structure, and weather conditions into consideration for accurate traffic prediction. Zhan *et al.* [27] propose an ensemble learning model that exploits the temporal characteristics of the data, and balances the accuracy of individual models and their mutual dependence through a covariance-regularizer. It combines predictions from multiple methods to generate a consensus traffic flow prediction. STDN *et al.* [28] treat each region and the corresponding neighbors as an image, exploiting local CNN to handle the spatial dependency. Furthermore, a long short term memory

network is adopted to capture long-term periodic temporal shifting. STDN is built on a regular structure. Zhang *et al.* [29] propose a multitask deep-learning framework that simultaneously predicts the node flow and edge flow throughout a spatio-temporal network. Zhou *et al.* [30] propose a wide-attention and deep-composite model. The wide-attention module can extract global key features from traffic flows via a linear model with self-attention mechanism. The deep-composite module can generalize local key features via convolutional neural network component and long short-term memory network component.

The traffic time series data are most irregular or non-euclidean data [3], [31]. The deep learning methods, e.g., convolutional networks, LSTM, can only process regular data. Therefore, the deep learning model requires to process the traffic data into regular data (series data or image-like data). However, GCN, on the other hand, with its innate ability to handle irregular or non-euclidean data, has been widely applied in traffic forecasting, which directly learn from the irregular traffic data.

### B. GCN Based Prediction

Li *et al.* [5] model the spatial dependency of traffic as a diffusion process and propose DCRNN. It captures the spatial dependency using bidirectional random walks on the graph, and the temporal dependency using the encoder-decoder architecture with scheduled sampling. Similarly, Chen *et al.* [32] exploit GCN for capturing spatial correlation and RNN for learning temporal dependencies. Attention mechanism is adopted by [4], [33] to assist capture spatial-temporal correlations. However, RNN based approaches suffer from gradient explosion/vanishing and computational inefficiency for capturing long-range temporal dependencies. Therefore, RNN is replaced with temporal convolution network [13], gated CNN [34], [35] and P3D [15] for capturing temporal correlation. Song *et al.* [18] propose a novel model STSGCN to capture spatial temporal correlations simultaneously.

Diao *et al.* [36] exploit GCN for spatial-temporal prediction, by replacing the Laplacian matrix with a dynamic Laplacian matrix estimator for capturing dynamics. They incorporate tensor decomposition into the deep learning framework, to enable timely learning with a low complexity. Similarly, Wu *et al.* [13] propose a self-adaptive adjacency matrix to discover the spatial dependencies in an end-to-end fashion, enhancing pre-defined spatial dependencies. Guo *et al.* [37] construct the dynamic road network graph matrices adaptively, based on spatial-temporal features extracted by a latent network. Bai *et al.* [16] infer the hidden inter-dependencies from data automatically by learning a node embedding dictionaries, without the need of pre-defined adjacency matrix. They propose two adaptive modules to enhance GCN: 1) a node adaptive parameter learning module to capture node-specific patterns; 2) a data adaptive graph generation module to infer the inter-dependencies among different traffic series automatically.

In this paper, we further exploit the fine-grained road topology to assist in learning traffic flow dynamics. We incorporate both the multi-scale static features and dynamic features to model the constantly changing spatial-temporal correlations. By selecting the top $l$ most correlated nodes, we exploit a residual 3D network for capturing the spatial and temporal dependencies synchronously.

## III. Problem Formulation

In this section, we define some notations used in this paper and formally define the setup of our problem.

*Definition 1 (Road Network): In a city, the road network is represented as a directed graph denoted as* $\mathcal{R} = (\mathcal{N}^r, \mathcal{E}^r)$, *where* $\mathcal{N}_r \triangleq \{N_1, N_2, \ldots, N_{|\mathcal{N}^r|}\}$ *means the set of road segments, and* $\mathcal{E}^r$ *stands for the set of edges. If* $e^r_{i,j} \triangleq (N_i, N_j) \in \mathcal{E}^r$, $N_i$ *and* $N_j$ *are directly connected.*

The graph $\mathcal{R}$ can also be described by an asymmetric adjacency matrix $\boldsymbol{A}^r \in \{0, 1\}^{|\mathcal{N}^r| \times |\mathcal{N}^r|}$, where $\boldsymbol{A}^r[i, j] = 1$ means $e^r_{i,j} \in \mathcal{E}^r$. For each road segment $N_i \in \mathcal{N}_r$, its fine-grained road attributes are represented as $\boldsymbol{Z}^r[i]$, which includes the location (latitude and longitude), type (motorway, trunk way, primary way, and secondary way[1]), length, number of lanes, speed limits and direction (the angle from the horizontal line to the road) of road segment $N_i$.

*Definition 2 (Traffic Network): Based on the road network* $\mathcal{R}$, *a sensor traffic network is represented as* $\mathcal{G} = (\mathcal{N}^s, \mathcal{E}^s)$, *where* $\mathcal{N}^s \triangleq \{v_1, v_2, \ldots, v_{|\mathcal{N}^s|}\}$ *is a set of* $|\mathcal{N}^s|$ *traffic sensor nodes (midpoints are added into the road network to make sure that each road segment has at most a sensor). Each sensor node can be mapped to a road segment, i.e.,* $\mathcal{N}^s \subseteq \mathcal{N}^r$.

For the sensor node set $\mathcal{N}^s$, its traffic information (e.g., traffic in-out flow, speed, capacity) at $t$-th time step is denoted as $\boldsymbol{Z}^s_t$. In this paper, only speed is considered due to the constraints of available dataset and our proposed model can be applied for prediction of other traffic information. Then, the traffic spatial-temporal prediction can be formally stated as:

### A. Spatial-Temporal Prediction Formulation

Given the road network $\mathcal{R} = (\mathcal{N}^r, \mathcal{E}^r)$, traffic network $\mathcal{G} = (\mathcal{N}^s, \mathcal{E}^s)$, and history $p$ steps' observations $\boldsymbol{X}^t = \{\boldsymbol{Z}^s_{t-p+1}, \boldsymbol{Z}^s_{t-p+2}, \ldots, \boldsymbol{Z}^s_t\}$, we aim to build a model $f_\theta$, which can accurately predict next $q$ time steps' traffic $\boldsymbol{Y}^t = \{\boldsymbol{Z}^s_{t+1}, \boldsymbol{Z}^s_{t+2}, \ldots, \boldsymbol{Z}^s_{t+q}\}$. The prediction model $f_\theta$ can be represented as:

$$[\boldsymbol{X}^t, \mathcal{R}, \mathcal{G}] \xrightarrow{f_\theta} \boldsymbol{Y}^t. \tag{1}$$

## IV. Methodology

### A. Static and Dynamic Feature Extractions

The correlations between nodes can be influenced by both the road networks and the real-time traffic information, which we summarized as static and dynamic factors, respectively. Two modules, i.e., multi-scale static feature extractor and dynamic feature encoder, are carefully orchestrated to extracted these two kinds of features, which are introduced as follows. The network architecture is shown in Fig.2.

[1]The fine-grained road attributes have been defined by OpenStreetMap, and please refer to https://wiki.openstreetmap.org/wiki/Map_Features

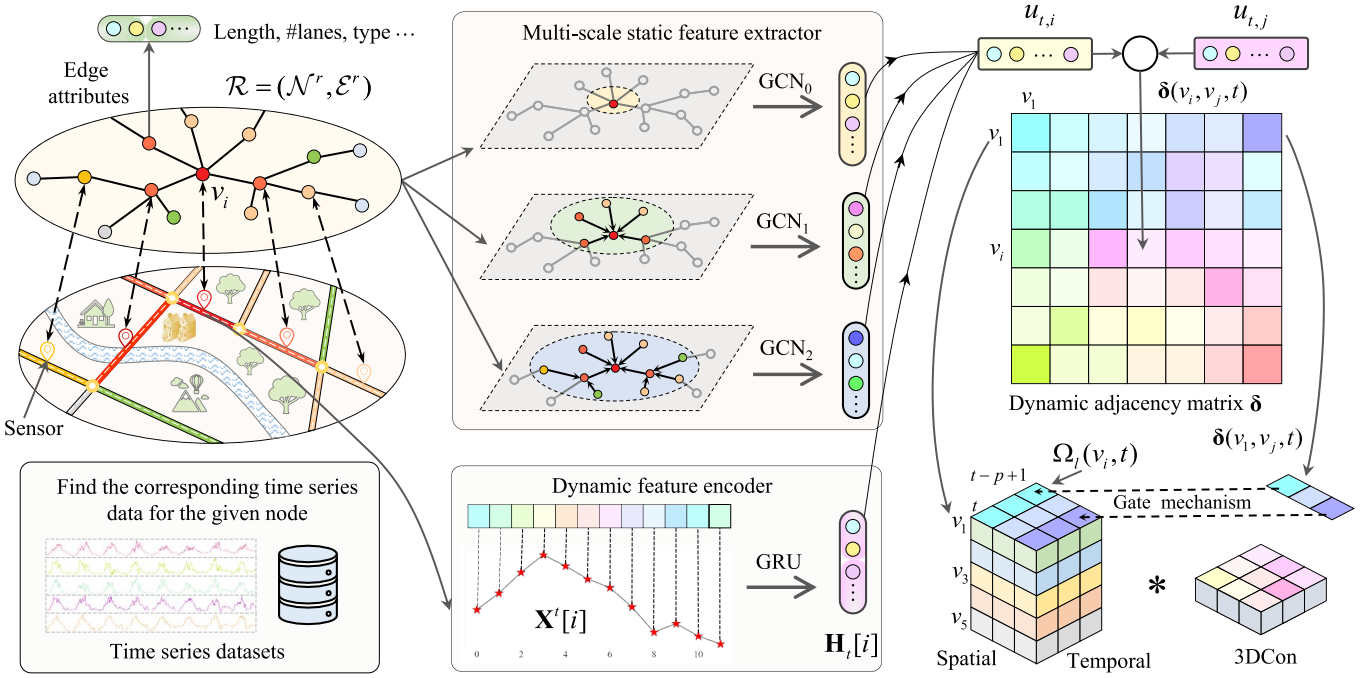



Fig. 2. Network architecture of ESTNet. First, GCN networks with multiple ranges are adopted to extract multi-scale static features and a GRU encoder is exploited to extract dynamic features. Then, these two features are fused as embeddings to determine dynamic correlations both in spatial and temporal dimensions. Consequently, at each time step $t$ for each node $v_i$, it is possible to dynamically determine a set of most correlated nodes $\Omega_l(v_i,t)$. Finally, a 3D convolution unit is applied to simultaneously capture the spatial-temporal dependencies.

*1) Multi-Scale Static Feature Extractor:* The road network with fine-grained road attributes can be applied to extract the static features. The fine-gained road attributes of two individual nodes, i.e., $\boldsymbol{Z}^r[i]$ and $\boldsymbol{Z}^r[j]$, can represent the similarity of these two nodes. For example, two nodes with the same type (i.e., both are highway) and the same number of lanes share a more similar feature. The neighboring nodes indicate structural equivalence or second-order proximity [38], which can be interpreted as nodes with shared neighbors being likely to be similar. The larger range of surrounding nodes presents the functionality of the area (transportation hub or residential district) [39], i.e., similar road topology distribution implies a higher affinity. Therefore, we should consider a different range of nodes to extract multi-scale features to fully represent the static correlations among nodes.

We apply GCN proposed by [14] for efficient multi-scale feature extraction, and the propagation rule is denoted as follows:

$$\boldsymbol{H}=\sigma(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{W}), \tag{2}$$

where $\boldsymbol{X}$ and $\boldsymbol{H}$ are the input and output graph signals, $\tilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}$ is the adjacency matrix with added self-connections, $\tilde{\boldsymbol{D}}=diag(\tilde{\boldsymbol{A}}\boldsymbol{1})$ is degree matrix of $\tilde{\boldsymbol{A}}$, $\boldsymbol{W}$ is the trainable weight matrix, and $\sigma$ is activation function.

Suppose the maximum range of nodes considered is $K$ hops. The feature extraction network for $k$-th hop range can be denoted as GCN$_k$, $0\le k\le K$. Therefore, if $k=0$, only features of individual nodes are considered, and GCN$_0$ can be denoted as

$$\text{GCN}_0(\boldsymbol{X})=\sigma(\boldsymbol{I}\boldsymbol{X}\boldsymbol{W}). \tag{3}$$

Inspired by diffusion process [5] and N-GCN [40], if $k>0$, GCN$_k$ can be denoted as

$$\text{GCN}_k(\boldsymbol{X})=\sigma\big((\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}})^k\boldsymbol{X}\boldsymbol{W}\big). \tag{4}$$

Therefore, given the adjacency matrix $\boldsymbol{A}^r$ and graph signal $\boldsymbol{Z}^r$ of the road network $\mathcal{R}$, the extracted multi-scale static features [5], [13] are denoted as $\{\text{GCN}_0(\boldsymbol{Z}^r),\text{GCN}_1(\boldsymbol{Z}^r),\ldots,\text{GCN}_K(\boldsymbol{Z}^r)\}$, with GCN$_k(\boldsymbol{Z}^r)$ defined as:

$$\text{GCN}_k(\boldsymbol{Z}^r)=\begin{cases}\sigma(\boldsymbol{I}\boldsymbol{Z}^r\boldsymbol{W}_0), & k=0,\\ \sigma((\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}^r\tilde{\boldsymbol{D}}^{-\frac{1}{2}})^k\boldsymbol{Z}^r\boldsymbol{W}_k), & \text{Otherwise.}\end{cases} \tag{5}$$

Note that we exploit separate networks with different trainable variables to extract different scale of features, i.e., $\boldsymbol{W}_i\neq\boldsymbol{W}_j$, if $i\neq j$ and $0\le i,j\le K$.

*2) Dynamic Feature Encoder:* For a node in the traffic network, i.e., $v_i\in\mathcal{G}$, the history observation $\boldsymbol{X}^t[i]$ contains the dynamic features for node $v_i$ at $t$-th time step. Inspired by seq2seq proposed by [41], we apply a gated recurrent unit (GRU) network to encode the sequence to a vector in a latent space, acting as the dynamic features [42]. The encoder can be denoted as:

$$\begin{aligned}z_t&=\sigma(\boldsymbol{W}_z\boldsymbol{X}^t[i]+\boldsymbol{U}_z\boldsymbol{H}_{t-1}[i]),\\ r_t&=\sigma(\boldsymbol{W}_r\boldsymbol{X}^t[i]+\boldsymbol{U}_r\boldsymbol{H}_{t-1}[i]),\\ \tilde{\boldsymbol{H}}_t[i]&=tanh(\boldsymbol{W}_h\boldsymbol{X}^t[i]+r_t\odot(\boldsymbol{U}_h\boldsymbol{H}_{t-1}[i])),\\ \boldsymbol{H}_t[i]&=(1-z_t)\odot\tilde{\boldsymbol{H}}_t[i]+z_t\odot\boldsymbol{H}_{t-1}[i],\end{aligned} \tag{6}$$

where $\odot$ denotes the element-wise multiplication, $\boldsymbol{W}_z,\boldsymbol{U}_z,\boldsymbol{W}_r,\boldsymbol{U}_r,\boldsymbol{W}_h,\boldsymbol{U}_h$ are the trainable parameters,

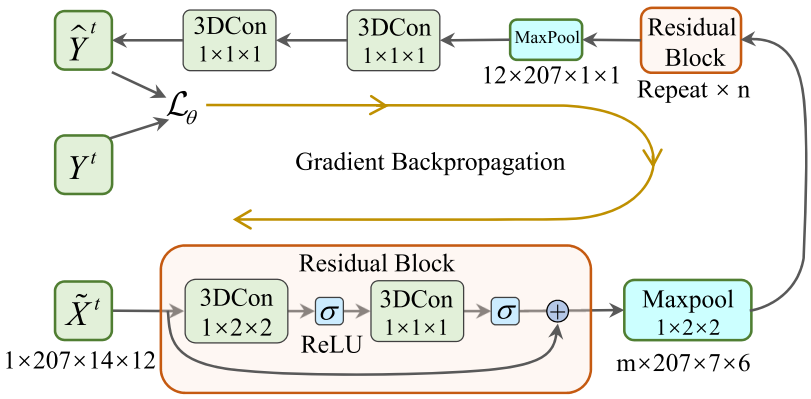



Fig. 3. Network structure of ESTNet on METR-LA.

and $\boldsymbol{H}_t[i]$ is the output, which also serves as the extracted dynamic features of node $v_i$ at $t$-th time step.

For each node in the traffic network, $v_i \in \mathcal{G}$, we concatenate its multi-scale static features and dynamic features as

$$\boldsymbol{u}_{t,i} = [\mathrm{GCN}_0(\boldsymbol{Z}^r)[i], \ldots, \mathrm{GCN}_K(\boldsymbol{Z}^r)[i], \boldsymbol{H}_t[i]]. \quad (7)$$

$\boldsymbol{u}_{t,i}$ can also be treated as the low dimensional embedding vector for $v_i$ at $t$-th time step.

### *B. Three Dimensional Graph Embedding Convolution Networks*

Given embeddings $\boldsymbol{u}_{t,i}$ and $\boldsymbol{u}_{t,j}$ of two nodes $v_i, v_j \in \mathcal{G}$, we exploit cosine similarity to define the dynamic correlations $\boldsymbol{\delta}(v_i, v_j, t)$ between $\boldsymbol{u}_{t,i}$ and $\boldsymbol{u}_{t,j}$. $\boldsymbol{\delta}(v_i, v_j, t)$ can be calculated by

$$\boldsymbol{\delta}(v_i, v_j, t) = \frac{\boldsymbol{u}_{t,i}.\boldsymbol{u}_{t,j}}{\|\boldsymbol{u}_{t,i}\| \ \|\boldsymbol{u}_{t,j}\|}. \quad (8)$$

Note that $\boldsymbol{\delta}$ reflects both the dynamic and static features, indicating the correlations among nodes in the traffic network $\mathcal{G}$.

The cosine similarity can be applied to define the adjacency matrix $\boldsymbol{A}^s$ of the road network $\mathcal{G}$, i.e., $\boldsymbol{A}^s[i,j] = \frac{1}{p}\sum_t \delta(v_i, v_j, t)$. Then, we can exploit several existing approaches, e.g., DCRNN, GMAN, SLC, etc., to capture the spatial-temporal correlations.

At $t$-th time step, we aim to predict $\boldsymbol{Y}^t \in \mathbb{R}^{C\times|\mathcal{N}^r|\times q}$ based on historical observations $\boldsymbol{X}^t \in \mathbb{R}^{C\times|\mathcal{N}^r|\times p}$, where $C$, $|\mathcal{N}^r|$, and $p$ are the number of input features and sensors, and the length of historical observations, respectively. Here, for each node $v_i \in \mathcal{G}$ at $t$-th time step, we select the top $l$ nodes with highest similarity as the most correlated nodes to $v_i$, which are denoted as $\Omega_l(v_i, t)$.

Since we consider that each node $v_i \in \mathcal{N}^s$ is correlated with $\Omega_l(v_i, t)$, we extend $\boldsymbol{X}^t$ to $\tilde{\boldsymbol{X}}^t \in \mathbb{R}^{C\times|\mathcal{N}^r|\times l\times p}$ by concatenating the graph signals of $\Omega_l(v_i, t)$ for each node $v_i$. Furthermore, a gating mechanism is adopted to further connect the similarity with real-time traffic, since it is powerful to control information flow through layers [13]. Consequently, $\tilde{\boldsymbol{X}}^t$ can be denoted as:

$$\tilde{\boldsymbol{X}}^t[i] = [\boldsymbol{X}^t[j_1]\cdot\delta(v_i, v_{j_1}, t), \boldsymbol{X}^t[j_2]\cdot\delta(v_i, v_{j_2}, t), \ldots, \boldsymbol{X}^t[j_l]\cdot\delta(v_i, v_{j_l}, t)], \ \forall v_{j_1}, v_{j_2}, v_{j_l} \in \Omega_l(v_i, t). \quad (9)$$

Therefore, the prediction problem can be expressed as:

$$\tilde{\boldsymbol{X}}^t \in \mathbb{R}^{C\times|\mathcal{N}^r|\times l\times p} \xrightarrow{f_\theta} \boldsymbol{Y}^t \in \mathbb{R}^{C\times|\mathcal{N}^r|\times q}. \quad (10)$$

To this end, inspired by $3D$ convolution, we exploit a three dimensional convolution unit (3DCon) to capture the spatial and temporal dependences synchronously, which is denoted as:

$$\boldsymbol{Y}^t[:,x,y] = \sigma\Big(\sum_{k,m,n} \tilde{\boldsymbol{X}}^t[:, x+k, o+m, y+n] \times \boldsymbol{W}[k,m,n] + b\Big), \quad (11)$$

where $\boldsymbol{W}$ is trainable variables and $\sigma(\cdot)$ is activation function.

### *C. End to End Embedding Spatial-Temporal Network*

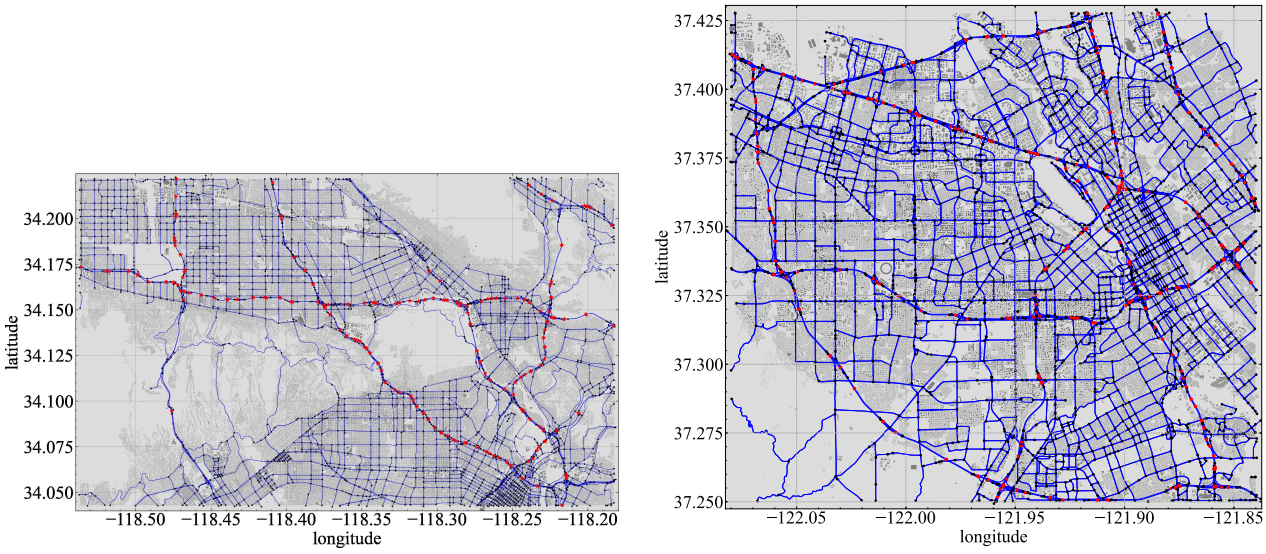
In order to capture the complicated and non-linear correlations, we should stack multiple 3DCon to construct a deeper network. Inspired by the remarkable success of ResNet networks on computer vision [43], we adopt a similar idea to construct ESTNet, as shown in Fig.3, based on the 3DCon unit. The basic block for ESTNet is the ResNet block, which consists of two 3DCon and two activations, with a residual connection. Then, a 3D maxpooling is applied, as shown in Fig.3. By stacking multiple residual blocks, ESTNet is able to capture the nonlinear and complex spatial-temporal correlations. Given $\tilde{\boldsymbol{X}}^t$ and $\boldsymbol{Y}^t$, ESTNet predicts $\hat{\boldsymbol{Y}}^t$, and the loss $\mathcal{L}_\theta$ is calculated as:

$$\mathcal{L}_\theta = \sum |\hat{\boldsymbol{Y}}^t - \boldsymbol{Y}^t|. \quad (12)$$

## V. Experiment

### *A. Datasets*

The effectiveness and strengths of ESTNet are validated on three publicly-available traffic datasets: METR-LA, and PEMS-BAY [5]. METR-LA contains 4 months (Mar 1st,




(a) METR-LA (b) PEMS-BAY

Fig. 4. The road network $\mathcal{R}$ and traffic network $\mathcal{G}$ of METR-LA, and PEMS-BAY, where dark circles mean intersections, red pentagrams stand for sensors and blue lines are road segments.

TABLE I
DETAIL INFORMATION OF ROAD NETWORK

| Datasets | METR-LA | PEMS-BAY |
|---|---|---|
| Location | Los Angeles | Bay Area |
| Number of sensors | 207 | 325 |
| Number of road segments | 6585 | 5966 |
| Total road length (km) | 3323.48 | 2825.73 |
| Number of intersections | 2367 | 2304 |
| Area ($km^2$) | 575.35 | 385.58 |
| Date | Mar 1st, 2012 to Jun 30th, 2012 | |
| Road attributes | (Length, type, speed limit, direction, lanes) | |

2012 to Jun 30th, 2012) of traffic speed data on 207 sensors located on the highway of Los Angeles County. PEMS-BAY contains four months (Jan 1st, 2017 to May 31st, 2017) of traffic speed data on 325 sensors in the Bay area. METR-LA and PEMS-BAY are traffic speed data. The data is aggregated to 5 minutes, which means there are 12 points in the data for each hour. We use traffic data from the past hour to predict the data for the next hour.

For each dataset, the corresponding road network is collected from OpenStreet Maps. Road networks are described by intersections, road segments (which are connections between intersections), and the fine-grained attributes of road segments, which is shown in Fig.4. Detailed information and comparison between these two datasets are shown in Tab.I. The distribution of sensors and the road network are shown in Fig.4, where dark circles mean intersections, red pentagrams stand for sensors and blue lines are road segments. As shown in Tab.I, METR-LA and PEMS-BAY has 207 and 305 sensors, respectively.

*1) Evaluation Metrics:* We adopt three well-known and widely-applied metrics to measure the performance of prediction. Let $\boldsymbol{Y} \in \mathbb{R}^{q \times \mathcal{N}_r \times C}$ be the ground truth traffic of all nodes, and $\tilde{\boldsymbol{Y}} \in \mathbb{R}^{q \times \mathcal{N}_r \times C}$ be the predicted values. The metrics are defined as follows:

1) Mean Absolute Error (MAE)

$$MAE = \frac{1}{q\mathcal{N}_r C} \sum_{t=1}^{q} \sum_{i=1}^{\mathcal{N}_r} \sum_{j=1}^{C} |\boldsymbol{Y}[t,i,j] - \tilde{\boldsymbol{Y}}[t,i,j]|. \tag{13}$$

2) Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{q\mathcal{N}_r C} \sum_{t=1}^{q} \sum_{i=1}^{\mathcal{N}_r} \sum_{j=1}^{C} (\boldsymbol{Y}[t,i,j] - \tilde{\boldsymbol{Y}}[t,i,j])^2}. \tag{14}$$

3) Mean Absolute qercentage Error (MAqE)

$$MAE = \frac{1}{q\mathcal{N}_r C} \sum_{t=1}^{q} \sum_{i=1}^{\mathcal{N}_r} \sum_{j=1}^{C} \left| \frac{\boldsymbol{Y}[t,i,j] - \tilde{\boldsymbol{Y}}[t,i,j]}{\boldsymbol{Y}[t,i,j]} \right|. \tag{15}$$

## B. Experimental Settings

All the speed readings are aggregated into 5 min' window, and we exploit historical observations of one hour to predict future one hour's traffic, i.e., $p = q = 12$.

*1) Structures of ESTNet and Learn Strategy:* We extract three scales ($K = 2$) of features from the road network and each extractor consists of three GCN layers. The GRU encoder consists of two-layer GRU. ReLU is selected as the non-linear activation function. The experiments are conducted on the Pytorch platform using an Intel(R) Xeon(R) Gold 5120, 2.20-GHz CPU, 128-GB RAM, and a GeForce GTX 2080-Ti 11G GPU. The optimizer is the Adam algorithm and the learning rate is set to 0.001.

TABLE II
PERFORMANCE OF DIFFERENT MODELS ON METR-LA DATASET, RESULTS WITH _ ARE THE BEST PERFORMANCE ACHIEVED BY BASELINES. (SMALLER VALUE MEANS BETTER PERFORMANCE)

| Model | | METR-LA (15 / 30 / 60 min) | | |
|---|---|---|---|---|
| | | MAE | RMSE | MAPE(%) |
| Traditional Approaches | VAR | 4.42 / 5.41 / 6.52 | 7.89 / 9.13 / 10.11 | 10.2 / 12.7 / 15.8 |
| | ARIMA | 3.99 / 5.15 / 6.90 | 8.21 / 10.45 / 12.23 | 9.6 / 12.7 / 17.4 |
| Deep learning methods | FNN | 3.99 / 4.23 / 4.49 | 8.17 / 7.94 / 8.69 | 9.9 / 12.9 / 14.0 |
| | FC-LSTM | 3.44 / 3.77 / 4.37 | 6.30 / 7.23 / 8.96 | 9.6 / 10.9 / 13.2 |
| GCN models | STGCN | 2.88 / 3.47 / 4.59 | 5.74 / 7.24 / 9.40 | 7.6 / 9.6 / 12.7 |
| | DCRNN | 2.77 / 3.15 / 3.60 | 5.38 / 6.45 / 7.59 | 7.3 / 8.8 / 10.5 |
| | STSGCN | 2.61 / 2.99 / 3.43 | 5.04 / 6.11 / 7.15 | 6.8 / 8.21 / 9.7 |
| | Graph WaveNet | 2.69 / 3.07 / 3.53 | 5.15 / 6.22 / 7.37 | 6.9 / 8.37 / 10.1 |
| | MRA-BGCN | 2.67 / 3.06 / 3.49 | 5.12 / 6.17 / 7.30 | 6.8 / 8.3 / 10.0 |
| | GMAN | 2.77 / 3.07 / 3.40 | 5.48 / 6.34 / 7.21 | 7.3 / 8.4 / 9.7 |
| | STGNN | 2.62 / 2.98 / 3.49 | *4.99* / *5.88* / *6.94* | *6.6* / *7.8* / *9.7* |
| | SLCNN | *2.53* / *2.88* / *3.30* | 5.18 / 6.15 / 7.20 | 6.7 / 8.0 / *9.7* |
| Our proposed approach | ESTNet | **2.32 / 2.71 / 3.18** | **4.30 / 5.36 / 6.75** | **5.6 / 7.06 / 9.17** |
| | Improvements | 8.30% / 5.90% / 3.64% | 13.83% / 8.84% / 2.74% | 15.15% / 9.49% / 5.46% |

*2) Baselines:* ESTNet is compared against traditional machine learning, deep learning, and GCN models, which are introduced as follows:

1) **VAR**: a time series model that captures spatial correlations among all traffic series;
2) **ARIMA**: auto-regressive integrated moving average model;
3) **FNN**: feed forward neural network;
4) **FC-LSTM**: fully-connected long short-term memory (LSTM) network;
5) **STGCN** [34]: capturing spatial and temporal correlations by GCN and temporal convolutional networks, respectively;
6) **DCRNN** [5]: diffusion convolution recurrent neural network, which formulates the graph convolution with the diffusion process and combines GCN with recurrent models in an encoder-decoder manner for multi-step prediction;
7) **Graph WaveNet** [13]: it enhances GCN with a novel adaptive dependency matrix and exploit a stacked dilated 1D convolution for learning temporal correlations;
8) **MRA-BGCN** [44]: multi-range attentive bicomponent GCN;
9) **GMAN** [4]: graph multi-attention network for traffic prediction;
10) **STGNN** [45]: it enhances GCN with position-wise attention mechanism and exploits gated recurrent neural network with transformer layer to learn temporal correlations;
11) **SLCNN** [15]: it proposes structure learning convolution and pseudo three dimensional convolution for capturing temporal correlations.
12) **STSGCN** [18]: it synchronously models spatial-temporal correlations.

All methods apply the setting and hyper-parameters reported by the corresponding papers.

### C. Experimental Results

*1) Comparison of Models:* Tab.II and Tab.III demonstrate the results of ESTNet as compared with 12 baselines on METR-LA, and PEMS-BAY datasets, respectively. After a careful analysis of the results shown in these two tables, we can make the following observations:

- Traditional methods achieve a poor prediction performance since they do not consider the traffic conditions of the immediate past and the spatial-temporal correlations. Therefore, they are not sufficient to handle the dynamic changes or spatial dependencies of traffic flow. As the data sequence is generated by sliding a window from original traffic data, the result of HA is invariant to the increases in the forecasting horizon.
- Deep learning based models (FNN, and FC-LSTM) perform better than traditional methods since they can model more complex and non-linear traffic flows. FNN needs to mess up the spatial-temporal data to form a long series, which breaks the spatial-temporal dependencies and makes it harder to capture the correlations. Consequently, the temporal neural network based methods FC-LSTM achieve a comparatively better performance than FNN.
- GCN based models (STGCN, DCRNN, Graph WaveNet, MRA-BGCN, GMAN, STGNN, SLCNN, and our proposed ESTNet) perform better than traditional methods and deep learning methods, since they model the sensors as a graph and could handle the non-regular data structures. Among the GCN based models, STSGCN performs poorly. If multiple spatial graphs of different time steps are concatenated, a larger adjacency matrix is produced by STSGCN, and therefore it suffers from long-sequence temporal dependencies, which are hard to capture and approximate.
- Our proposed ESTNet outperforms all the baseline models in all settings. ESTNet adopts a novel degree-aware

TABLE III

PERFORMANCE OF DIFFERENT MODELS ON PEMS-BAY DATASET, RESULTS WITH _ ARE THE BEST PERFORMANCE ACHIEVED BY BASELINES. (SMALLER VALUE MEANS BETTER PERFORMANCE)

| Model | | PEMS-BAY (15 / 30 / 60 min) | | |
|---|---|---|---|---|
| | | MAE | RMSE | MAPE(%) |
| Traditional Approaches | VAR | 1.74 / 2.32 / 2.93 | 3.16 / 4.25 / 5.0 | 3.6 / 5.44 / 6.5 |
| | ARIMA | 1.62 / 2.33 / 3.38 | 3.3 / 4.76 / 6.50 | 3.5 / 5.4 / 8.3 |
| Deep learning methods | FNN | 2.2 / 2.3 / 2.46 | 4.42 / 4.63 / 4.98 | 5.2 / 5.4 / 5.9 |
| | FC-LSTM | 2.05 / 2.20 / 2.37 | 4.19 / 4.55 / 4.96 | 4.8 / 5.2 / 5.7 |
| GCN models | STGCN | 1.36 / 1.81 / 2.49 | 2.96 / 4.27 / 5.69 | 2.9 / 4.2 / 5.8 |
| | DCRNN | 1.38 / 1.74 / 2.07 | 2.95 / 3.97 / 4.74 | 2.9 / 3.9 / 4.9 |
| | STSGCN | 1.24 / 1.51 / 1.84 | 2.62 / 3.43 / 4.31 | 2.6 / 3.5 / 4.4 |
| | Graph WaveNet | 1.30 / 1.63 / 1.95 | 2.74 / 3.70 / 4.52 | 2.7 / 3.7 / 4.6 |
| | MRA-BGCN | 1.29 / 1.61 / 1.91 | 2.72 / 3.67 / 4.46 | 2.9 / 3.8 / 4.6 |
| | GMAN | 1.34 / 1.62 / 1.86 | 2.82 / 3.72 / 4.32 | 2.8 / 3.6 / 4.3 |
| | STGNN | *1.17* / *1.46* / *1.83* | *2.43* / *3.27* / *4.20* | *2.3* / *3.1* / *4.2* |
| | SLCNN | 1.44 / 1.72 / 2.03 | 2.90 / 3.81 / 4.53 | 3.0 / 3.9 / 4.8 |
| Our proposed approach | ESTNet | **1.16 / 1.34 / 1.72** | **2.34 / 2.85 / 3.88** | **2.23 / 2.75 / 3.8** |
| | Improvements | 0.85% / 8.22% / 6.01% | 3.70% / 12.84% / 7.62% | 3.04% / 11.29% / 9.52% |

TABLE IV

ABLATION STUDY OVER DIFFERENT FUSION METHODS FOR COMBINING THE SIMILARITY AND THE TRAFFIC, I.E., CONCAT, GATE, AND PLUS MECHANISM

| Methods | METR-LA (15 / 30 / 60 min) | | |
|---|---|---|---|
| | MAE | RMSE | MAPE(%) |
| Gate mechanism | 2.32 / 2.71 / 3.18 | 4.30 / 5.36 / 6.75 | 5.62 / 7.06 / 9.17 |
| Concat mechanism | 2.33 / 2.70 / 3.28 | 4.32 / 5.31 / 6.95 | 5.83 / 7.09 / 9.37 |
| Plus mechanism | 2.34 / 2.73 / 3.19 | 4.27 / 5.46 / 6.79 | 5.61 / 7.13 / 9.23 |

method to learn the essential influence among road attributes, intersections, and traffic flows. Therefore, it achieves the best performance, demonstrating the effectiveness and advantages of our proposed approach.

Our proposed model achieves the best performance with a significant margin. The last row of Tab.II indicates the improvements of ESTNet over the best model in the corresponding metric, e.g., the best MAE for 15 minutes of baselines on METR-LA is 2.53, and that of ESTNet is 2.32. Therefore, the relative improvement is $(2.53 - 2.32)/2.53 = 8.30\%$. As shown in Tab.III, ESTNet achieves tremendously better performance, e.g., with a 13.83% and 15.15% improvement for RMSE and MAPE metric for predicting future 15 minutes' traffic on METR-LA, respectively. Similar improvements can also be found on PEMS-BAY. ESTNet achieves an average 8.15% and 7.01% relative improvement to the existing best results on METR-LA and PEMS-BAY, respectively.

To better understand the model, we visualize forecasting results. Fig.7 shows the visualization of 15 minutes, 30 minutes, 45 minutes, and 1 hour ahead forecasting.

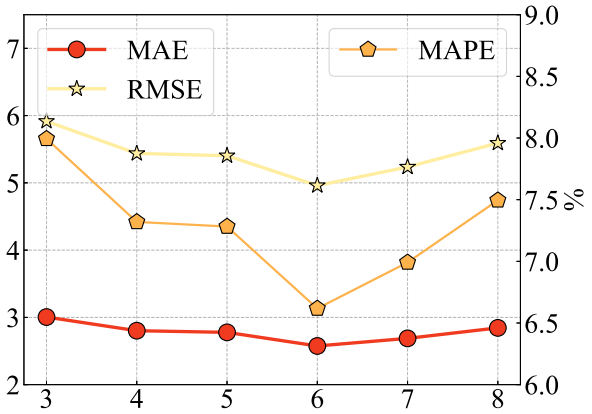
*2) Ablation Study:* To better evaluate the performance of ESTNet, we conduct a comprehensive ablation study. Tab.V shows the effectiveness of multi-scale static features and dynamic features. We can observe that: 1) Fusing the static and dynamic features outperforms the model with only partial features, which verify that both features are beneficial for determining the correlations. 2) Multi-scale static features help the traffic prediction by aggregating efficient features to distinguish the correlations. However, with the number of scales $K$ increase from 3 to 4, the performance is decreasing and ESTNet achieves the best performance when $K$ is 2 or 3.

Furthermore, extensive experiments have been conducted to evaluate the performance of ESTNet with a different number of ResNet blocks and different correlated neighbors $l$, which are shown in Fig.5(a) and Fig.5(b), respectively. ESTNet achieves the best performance with 6 ResNet blocks on METR-LA. With more ResNet blocks, the performance decrease due to over-fitting. ESTNet assumes that each node is correlated with top $l$ neighbors and Fig.5(b) shows the performance with different $l$. More correlated neighbors lead to higher performance in the beginning (when $l < 20$). However, when the number of correlated neighbors continues increase, the performance degrades. Meanwhile, larger $l$ causes heavier computation overhead. Therefore, to achieve a balance between accuracy and speed, we select $l = 14$ for our experiments.

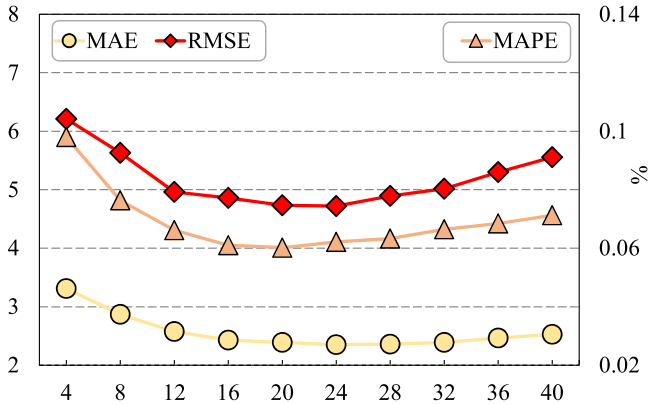
We apply a gate mechanism to connect the learned similarity and the real-time traffic in (9), since it is powerful to control information flow through layers [13]. To evaluate the performance of the gate mechanism, it is compared with the following fusion methods for combining the similarity and the traffic, which are introduced as follows:

TABLE V
ABLATION STUDY, WITH S AND D STAND FOR STATIC AND DYNAMIC FEATURES, RESPECTIVELY

| Methods | $K$ | METR-LA (15 / 30 / 60 min) | | |
|---|---|---|---|---|
| | | MAE | RMSE | MAPE(%) |
| D | _ | 2.84 / 3.40 / 4.30 | 5.59 / 6.97 / 8.87 | 7.23 / 9.40 / 12.85 |
| S | 2 | 3.31 / 4.10 / 5.25 | 6.43 / 8.11 / 10.24 | 8.84 / 11.90 / 16.58 |
| S+D | 0 | 2.50 / 2.74 / 3.33 | 4.74 / 5.42 / 6.85 | 6.14 / 7.16 / 9.31 |
| | 1 | 2.47 / 2.73 / 3.36 | 4.68 / 5.46 / 6.97 | 6.07 / 7.20 / 9.54 |
| | 2 | **2.32** / 2.71 / **3.18** | **4.30** / 5.36 / 6.75 | **5.60** / 7.06 / 9.17 |
| | 3 | 2.45 / **2.68** / 3.22 | 4.62 / **5.29** / **6.64** | 5.98 / **6.93** / **8.98** |
| | 4 | 2.47 / 2.71 / 3.27 | 4.67 / 5.36 / 6.74 | 6.07 / 7.06 / 9.17 |




(a) ResNet blocks




(b) $l$

Fig. 5. Performance with different network layers and $l$ on METR-LA.

1) **Concat mechanism:** This method connect the similarity and traffic through concatenation. Therefore, (9) can be formally defined as:

$$\tilde{\boldsymbol{X}}^t[i] = [[\boldsymbol{X}^t[j_1], \delta(v_i, v_{j_1}, t)], [\boldsymbol{X}^t[j_2], \delta(v_i, v_{j_2}, t)], \ldots, [\boldsymbol{X}^t[j_l], \delta(v_i, v_{j_l}, t)]], \quad \forall\, v_{j_1}, v_{j_2}, v_{j_l} \in \Omega_l(v_i, t). \tag{16}$$

2) **Plus mechanism:** This method connect the similarity and traffic through plus. Therefore, (9) can be formally defined as:

$$\tilde{\boldsymbol{X}}^t[i] = [[\boldsymbol{X}^t[j_1] + \delta(v_i, v_{j_1}, t)], [\boldsymbol{X}^t[j_2] + \delta(v_i, v_{j_2}, t)], \ldots, [\boldsymbol{X}^t[j_l] + \delta(v_i, v_{j_l}, t)]], \quad \forall\, v_{j_1}, v_{j_2}, v_{j_l} \in \Omega_l(v_i, t). \tag{17}$$

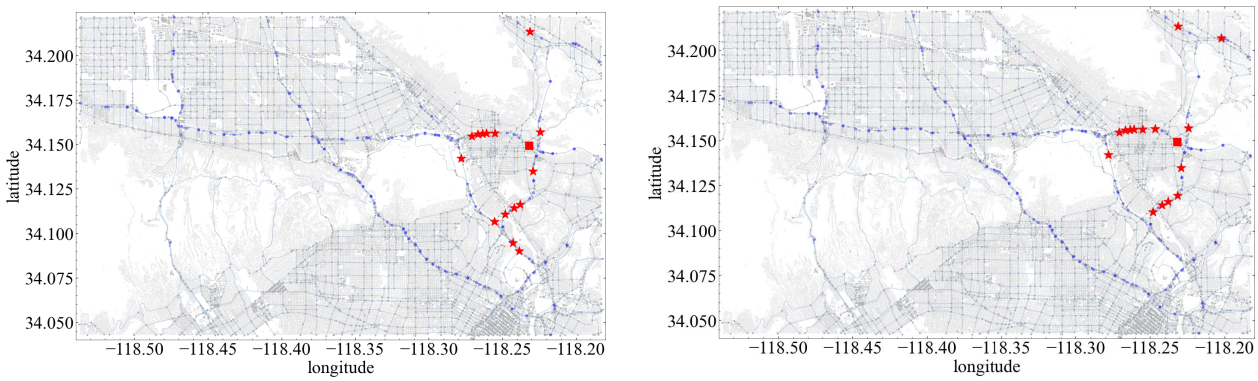
TABLE VI
COMPUTATION TIME EXPERIMENTS OF ESTNET AS COMPARED WITH DCRNN, ASTGCN, STGCN, AGCRN, AND GRAPH WAVENET

| Model | Training time (epoch) |
|---|---|
| ESTNet | 97.21 s |
| DCRNN | 213.24 s |
| Graph WaveNet | 45.78 s |
| STGCN | 95.87 s |
| ASTGCN | 289.87 s |
| AGCRN | 208.38 s |

The experiments are shown in Tab.IV, all three fusion mechanisms achieve similar performance. The gate mechanism achieves the best performance as for 15 minutes' traffic in terms of MAE while the concat mechanism achieves the best performance as for 30 minutes' traffic in terms of RMSE.

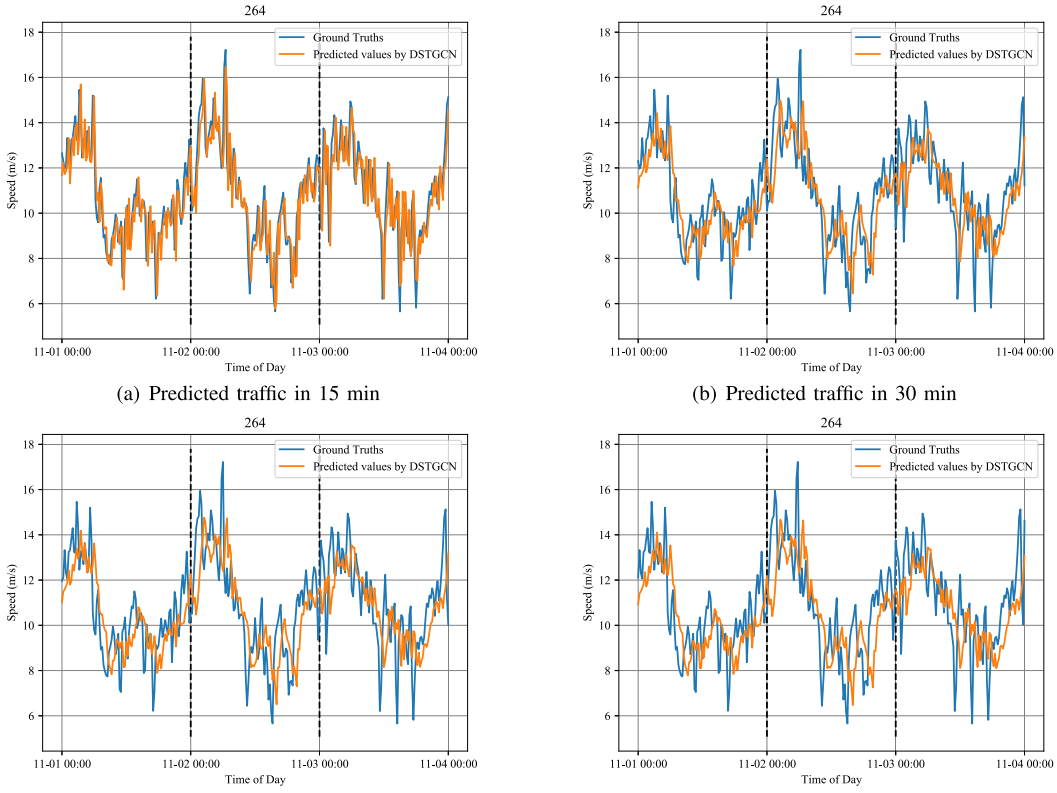
To further demonstrate the dynamic dependencies and the effectiveness of our proposed ESTNet, we visualize dynamic correlations $\boldsymbol{\delta}(v_i, v_j, t)$. For each node, we visualize the top 15 correlated nodes which has the maximum $\boldsymbol{\delta}(v_i, v_j, t)$. The results are shown in Fig.6, where the blue circles mean the sensor locations, red squares are the considered sensor and red pentagrams stand for top 15 correlated sensors. These two figures show that the correlations among the nodes are dynamic.

*3) Computation Time:* We compare the computation cost of ESTNet with DCRNN, ASTGCN, STGCN, AGCRN, and Graph WaveNet on the METR-LA dataset, which is shown in Tab.VI. ESTNet achieves 97.21 s/epoch for training, while that of DCRNN, ASTGCN, and AGCRN are 213.24, 289.87 and 208.38 s/epoch, respectively, and that of Graph WaveNet and STGCN are 45.78 and 95.87s/epoch, respectively. ESTNet is slower than Graph WaveNet and STGCN but faster than DCRNN, ASTGCN, and AGCRN. However, ESTNet achieves a significantly better performance than all the baseline methods.




(a) (b)

Fig. 6. Visualization of the correlated nodes at different time on METR-LA, where the blue circles mean the sensor locations, red squares are the considered sensor and red pentagrams stand for top 15 correlated sensors. These two figures show that the correlations among the nodes are dynamic.




(a) Predicted traffic in 15 min (b) Predicted traffic in 30 min

(c) Predicted traffic in 45 min (d) Predicted traffic in 60 min

Fig. 7. Prediction results of DSTGCN on an edge within three days.

## VI. Conclusion and Future Work

In this paper, we present a novel embedded spatial-temporal network, ESTNet, which extracts efficient features to dynamically capture the correlations between nodes, thus eliminating the requirements of a pre-defined adjacency matrix. A multi-range GCN network is adopted to extract multi-scale static features from a fine-grained road network and a GRU encoder is applied to extract dynamic features from real-time traffic. These two kinds of features are fused to determine the top correlated neighbors. Furthermore, a residual 3D network is applied to synchronously capture the spatial-temporal correlations. Experiments demonstrate that ESTNet outperforms the state-of-the-art with a significant margin. In the future, we will improve ESTNet for better computation efficiency.

## References

[1] G. Luo *et al.*, "Cooperative vehicular content distribution in edge computing assisted 5G-VANET," *China Commun.*, vol. 15, no. 7, pp. 1–17, Jul. 2018.

[2] G. Luo *et al.*, "Software-defined cooperative data sharing in edge computing assisted 5G-VANET," *IEEE Trans. Mobile Comput.*, vol. 20, no. 3, pp. 1212–1229, Mar. 2021.

[3] X. Yin, G. Wu, J. Wei, Y. Shen, H. Qi, and B. Yin, "Deep learning on traffic prediction: Methods, analysis and future directions," 2020, *arXiv:2004.08555*.

[4] C. Zheng, X. Fan, C. Wang, and J. Qi, "GMAN: A graph multi-attention network for traffic prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 1, pp. 1234–1241.

[5] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–16.

[6] H. Yao, Y. Liu, Y. Wei, X. Tang, and Z. Li, "Learning from multiple cities: A meta-learning approach for spatial-temporal prediction," in *Proc. World Wide Web Conf. (WWW)*, 2019, pp. 2181–2191.

[7] Q. Yuan, J. Li, H. Zhou, T. Lin, G. Luo, and X. Shen, "A joint service migration and mobility optimization approach for vehicular edge computing," *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 9041–9052, Aug. 2020.

[8] G. Luo, H. Zhang, H. He, J. Li, and F.-Y. Wang, "Multiagent adversarial collaborative learning via mean-field theory," *IEEE Trans. Cybern.*, vol. 51, no. 10, pp. 4994–5007, Oct. 2021.

[9] H. Zhang, G. Luo, J. Li, and F.-Y. Wang, "C2FDA: Coarse-to-fine domain adaptation for traffic object detection," *IEEE Trans. Intell. Transp. Syst.*, early access, Nov. 16, 2021, doi: 10.1109/TITS.2021.3115823.

[10] G. Luo, Q. Yuan, J. Li, S. Wang, and F. Yang, "Artificial intelligence powered mobile networks: From cognition to decision," 2022, *arXiv:2112.04263*.

[11] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 802–810.

[12] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 1655–1661.

[13] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, "Graph WaveNet for deep spatial-temporal graph modeling," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 1907–1913.

[14] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–14.

[15] Q. Zhang, J. Chang, G. Meng, S. Xiang, and C. Pan, "Spatio-temporal graph structure learning for traffic forecasting," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 1177–1185.

[16] L. Bai, L. Yao, C. Li, X. Wang, and C. Wang, "Adaptive graph convolutional recurrent network for traffic forecasting," in *Proc. NIPS*, vol. 33, 2020, pp. 1–12.

[17] Y. Zhang, Y. Li, X. Zhou, X. Kong, and J. Luo, "Curb-GAN: Conditional urban traffic estimation through spatio-temporal generative adversarial networks," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 842–852.

[18] C. Song, Y. Lin, S. Guo, and H. Wan, "Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 1, pp. 914–921.

[19] L. Wang, X. Geng, X. Ma, F. Liu, and Q. Yang, "Cross-city transfer learning for deep spatio-temporal prediction," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 1893–1899.

[20] J. Li *et al.*, "An end-to-end load balancer based on deep learning for vehicular network traffic control," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 953–966, Feb. 2019.

[21] H. Zhang, Y. Tian, K. Wang, W. Zhang, and F.-Y. Wang, "Mask SSD: An effective single-stage approach to object instance segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 2078–2093, 2020.

[22] H. Zhang, K. Wang, and F. Wang, "Advances and perspectives on applications of deep learning in visual object detection," *Acta Automatica Sinica*, vol. 43, no. 8, pp. 1289–1305, 2017.

[23] S. Li, S. Zhao, B. Cheng, and J. Chen, "Noise-aware framework for robust visual tracking," *IEEE Trans. Cybern.*, vol. 52, no. 2, pp. 1179–1192, Feb. 2022.

[24] D. Kang, Y. Lv, and Y.-Y. Chen, "Short-term traffic flow prediction with LSTM recurrent neural network," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2017, pp. 1–6.

[25] Y. Liu, H. Zheng, X. Feng, and Z. Chen, "Short-term traffic flow prediction with Conv-LSTM," in *Proc. 9th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2017, pp. 1–6.

[26] Z. Zheng, Y. Yang, J. Liu, H.-N. Dai, and Y. Zhang, "Deep and embedded learning approach for traffic flow prediction in urban informatics," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3927–3939, Oct. 2019.

[27] H. Zhan, G. Gomes, X. S. Li, K. Madduri, A. Sim, and K. Wu, "Consensus ensemble system for traffic flow prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 12, pp. 3903–3914, Dec. 2018.

[28] H. Yao, X. Tang, H. Wei, G. Zheng, and Z. Li, "Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction," in *Proc. Conf. Artif. Intell.*, vol. 33, 2019, pp. 5668–5675.

[29] J. Zhang, Y. Zheng, J. Sun, and D. Qi, "Flow prediction in spatio-temporal networks based on multitask deep learning," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 3, pp. 468–478, Mar. 2019.

[30] J. Zhou, H.-N. Dai, H. Wang, and T. Wang, "Wide-attention and deep-composite model for traffic flow prediction in transportation cyber–physical systems," *IEEE Trans. Ind. Informat.*, vol. 17, no. 5, pp. 3431–3440, May 2021.

[31] D. A. Tedjopurnomo, Z. Bao, B. Zheng, F. 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, Apr. 2020.

[32] C. Chen, K. Li, S. G. Teo, X. Zou, and Z. Zeng, "Gated residual recurrent graph neural networks for traffic prediction," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 485–492.

[33] R. Huang, C. Huang, Y. Liu, G. Dai, and W. Kong, "LSGCN: Long short-term traffic prediction with graph convolutional networks," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, vol. 7, Jul. 2020, pp. 2355–2361.

[34] 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. Artif. Intell.*, Jul. 2018, pp. 3634–3640.

[35] R. Dai, S. Xu, Q. Gu, C. Ji, and K. Liu, "Hybrid spatio-temporal graph convolutional network: Improving traffic prediction with navigation data," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 3074–3082.

[36] Z. Diao, G. Wang, D. Zhang, Y. Liu, K. Xie, and S. He, "Dynamic spatial-temporal graph convolutional neural networks for traffic forecasting," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 890–897.

[37] K. Guo, Y. Hu, Z. Qian, Y. Sun, J. Gao, and B. Yin, "Dynamic graph convolution network for traffic forecasting based on latent network of Laplace matrix estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 2, pp. 1009–1018, Feb. 2022.

[38] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "LINE: Large-scale information network embedding," in *Proc. 24th Int. Conf. World Wide Web*, May 2015, pp. 1067–1077.

[39] H. Liu, Y. Tong, P. Zhang, X. Lu, J. Duan, and H. Xiong, "Hydra: A personalized and context-aware multi-modal transportation recommendation system," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 2314–2324.

[40] S. Abu-El-Haija, A. Kapoor, B. Perozzi, and J. Lee, "N-GCN: Multi-scale graph convolution for semi-supervised node classification," in *Proc. 35th Uncertainty Artif. Intell. Conf.* (Proceedings of Machine Learning Research), vol. 115. TelAviv, Israel, 2020, pp. 841–851. [Online]. Available: https://proceedings.mlr.press/v115/abu-el-haija20a.html

[41] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 27, 2014, pp. 3104–3112.

[42] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. NIPS Workshop Deep Learn.*, Dec. 2014, pp. 1–9.

[43] F. He, T. Liu, and D. Tao, "Why ResNet works? Residuals generalize," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5349–5362, Dec. 2020.

[44] W. Chen, L. Chen, Y. Xie, W. Cao, and X. Feng, "Multi-range attentive bicomponent graph convolutional network for traffic forecasting," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 4, pp. 3529–3536.

[45] X. Wang *et al.*, "Traffic flow prediction via spatial temporal graph neural network," in *Proc. Web Conf.*, Apr. 2020, pp. 1082–1092.


**Guiyang Luo** (Graduate Student Member, IEEE) received the Ph.D. degree in computer science and technology from the Beijing University of Posts and Telecommunications (BUPT), China, in 2020. He is currently a Post-Doctoral Fellow at the State Key Laboratory of Networking and Switching Technology, BUPT. His current research interests include multi-agent systems and intelligent transportation systems.

**Hui Zhang** received the B.S. degree in automation from the Beijing Jiaotong University, Beijing, China, in 2015, and the Ph.D. degree in control theory and control engineering from the University of Chinese Academy of Sciences (UCAS), Beijing, in 2020. From August 2018 to October 2019, she was supported by UCAS as a joint-supervision Ph.D. student with the University of Rhode Island, Kingston, USA. She is currently a Lecturer with the School of Computer and Information Technology, Beijing Jiaotong University. Her research interests include computer vision, pattern recognition, and intelligent transportation systems.

**Quan Yuan** (Member, IEEE) received the Ph.D. degree in computer science and technology from the Beijing University of Posts and Telecommunications (BUPT), China, in 2018. He is currently working as a Post-Doctoral Fellow at the State Key Laboratory of Networking and Switching Technology, BUPT. His current research interests include crowdsensing, connected vehicle, mobile internet, and intelligent transportation systems.

**Jinglin Li** (Member, IEEE) received the Ph.D. degree in computer science and technology from the Beijing University of Posts and Telecommunications in 2004. He is currently a Professor of computer science and technology and the Director of the Switching and Intelligent Control Research Center (SICRC), State Key Laboratory of Networking and Switching Technology, China. His research interests are mainly in the areas of mobile internet, the Internet of Things, the Internet of Vehicles, convergence networks, and service technologies.

**Fei-Yue Wang** (Fellow, IEEE) is currently the Director of the State Key Laboratory for Management and Control of Complex Systems. His current research focuses on methods and applications for parallel systems, social computing, and knowledge automation. He was the President of IEEE ITS Society (2005–2007), the Chinese Association for Science and Technology (CAST, USA) in 2005, and the American Zhu Kezhen Education Foundation (2007–2008), and the Vice President of the ACM China Council (2010–2011). Since 2008, he has been the Vice President and the Secretary General of the Chinese Association of Automation. He is an elected fellow of INCOSE, IFAC, ASME, and AAAS. In 2007, he received the 2nd Class National Prize in Natural Sciences of China and awarded the Outstanding Scientist by ACM for his work in intelligent control and social computing. He received the IEEE ITS Outstanding Application and Research Awards in 2009 and 2011 and the IEEE SMC Norbert Wiener Award in 2014. Since 1997, he has been serving as the General or Program Chair for more than 20 IEEE, INFORMS, ACM, ASME conferences. He was the Founding Editor-in-Chief of the *International Journal of Intelligent Control and Systems* (1995–2000), the Founding EiC of *IEEE Intelligent Transportation Systems Magazine* (2006–2007), and the EiC of IEEE INTELLIGENT SYSTEMS (2009–2012) and IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS (2009–2016). Currently, he is the EiC of *China's Journal of Command and Control*.