Hindawi
Journal of Advanced Transportation
Volume 2022, Article ID 2806183, 17 pages
https://doi.org/10.1155/2022/2806183

Research Article

ST-AGRNN: A Spatio-Temporal Attention-Gated Recurrent Neural Network for Traffic State Forecasting

Jian Yang[1,2] **Jinhong Li**[1,2] **Lu Wei**[1] **Lei Gao**[1,2] **and Fuqi Mao**[2]

[1]*Beijing Key Lab of Urban Road Traffic Intelligent Technology, North China University of Technology, Beijing 100144, China*
[2]*School of Computer Science and Technology, North China University of Technology, Beijing 100144, China*

Correspondence should be addressed to Jian Yang; yanj200045@163.com

Received 2 June 2022; Revised 26 July 2022; Accepted 13 September 2022; Published 3 October 2022

Academic Editor: Yanming Shen

Accurate traffic state prediction plays an important role in traffic guidance, travel planning, etc. Due to the existence of complex spatio-temporal relationships, there are some challenges in forecasting. Firstly, in terms of spatial correlation, some models only consider the road network structure information, and ignore the relative location relationships between nodes. Secondly, some models ignore the different impacts of nodes in the global road network on traffic. To solve these problems, we propose a new traffic state-forecasting model, namely, spatio-temporal attention-gated recurrent neural network (ST-AGRNN). In the proposed model, structure-based and location-based localized spatial features are obtained simultaneously by Graph Convolutional Networks (GCNs) and DeepWalk. The localized temporal features are obtained by gated recurrent unit (GRU). The attention-based approach is used to obtain global spatio-temporal features. Experimental validation is performed with two real-world public datasets, and the results show that the ST-AGRNN model outperforms the state-of-the-art methods.

1. Introduction

Traffic congestion is a common problem faced by almost all major cities. Because of traffic congestion, a lot of manpower and material resources are wasted every year. Accurate and real-time traffic state prediction is the basis to solve the problem of traffic congestion. On the one hand, people can plan their trips in advance through traffic-state information. On the other hand, traffic managers also conduct effective traffic guidance and management through traffic state prediction information. At the same time, traffic prediction is a typical spatio-temporal problem, and the inherent nonlinearity and complexity of traffic affect the accuracy of prediction. Therefore, integrated consideration of temporal and spatial characteristics is necessary for traffic state prediction.

Taking the spatio-temporal correlation in Figure 1 as an example, there are localized spatio-temporal correlations and global spatio-temporal correlations. Each node will have influence on the traffic of its neighbors because it is physically connected with its neighbors and belongs to the relationship between upstream and downstream, which is spatial dependence. At the same time, each node will also affect itself at the next time step, which is temporal dependence. These are localized spatio-temporal correlations. In addition, a busy intersection has influence on the traffic of the entire region, which is the global spatio-temporal correlation in the road network. Obtaining this correlation is crucial to spatio-temporal data prediction.

In previous studies, various deep learning approaches were used to model spatio-temporal correlations, including stacked autoencoders (SAEs) [1], recurrent neural networks (RNNs) [2], generative adversarial networks (GANs) [3], transformer [4, 5], convolutional neural networks (CNNs) [6], and Spatio-Temporal Graph Convolutional Networks (STGCN) [7]. The SAEs acquire spatial and temporal correlations through unsupervised learning. The RNNs extract temporal features through the gate mechanism. The GANs extract spatio-temporal features through generators and discriminators and the transformer model spatial and temporal dependencies through encoder–decoder architecture. The

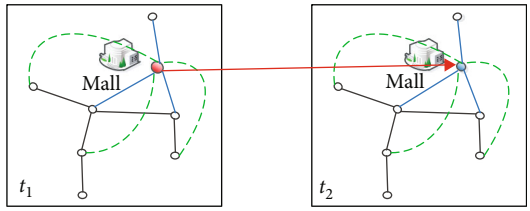


Figure 1: The influence of nodes in spatio-temporal correlations networks. The solid blue lines represent node spatio correlations. The red arrow represents the node temporal correlations. The green dash lines represent the global spatio-temporal correlations.

CNNs and GCNs obtain spatial features through convolution operation. However, these methods only capture localized spatio-temporal correlations.

Recently, attention mechanisms have received increasing attention. Because they are effective in identifying the relevance of inputs in prediction, components with high relevance are given greater attention. They are successfully applied in many fields, such as natural language processing (NLP) [8], computer vision (CV) [9, 10], and speech recognition [11]. Attention-based traffic forecasting has also developed rapidly in recent years. For example, attention temporal graph convolutional network (A3T-GCN) [12] uses attention mechanism to obtain global temporal and spatial correlations. However, it ignores location-based localized spatial information.

To obtain complex localized and global spatio-temporal correlations, we propose a novel deep learning architecture—spatio-temporal attention-gated recurrent neural network (ST-AGRNN)—for traffic state prediction. To fully exploit the localized spatio-temporal correlations, ST-AGRNN learns structure-aware graph embedding information through a GCN, and obtains position-aware information through DeepWalk. To tackle temporal dependencies, a gated recurrent unit (GRU) is used. Finally, in order to fully exploit the global spatio-temporal correlations, the attention mechanism is used to obtain spatio-temporal correlations about the networks.

The main contributions of this work are as follows:

(i) we propose a new localized spatial feature extraction method by combining DeepWalk with a GCN, where DeepWalk obtains position-aware information and the GCN obtains structure-aware graph embedding information

(ii) Traffic state is a time series data. The current traffic state will affect the traffic state at the next time step. GRU is used to obtain localized temporal correlation between traffic data

(iii) Attention mechanisms are introduced to obtain global spatio-temporal correlations about networks. Different nodes have different impacts on the traffic state, and the attention mechanism can obtain the weight of nodes from the historical traffic state, representing the global spatio-temporal correlations of network

(iv) Our experiments applying ST-AGRNN to traffic state prediction show that ST-AGRNN outperforms 12 state-of-the-art methods in terms of both accuracy and robustness on two benchmark datasets

2. Literature Review

2.1. Traffic State Forecasting. Time series data modeling and prediction are widely used in many fields [13, 14]. Traffic state data is a typical time series data. There are two main categories in traffic forecasting: statistical methods and machine learning methods. Statistical methods include autoregressive integrated moving average (ARIMA), the Kalman filter (KF), Markov chains, exponential smoothing (ES), and Bayesian networks. In the 1970s, Ahmed and Cook [15] used ARIMA to predict short-term traffic flow. Hamed et al. [16] later applied a simple ARIMA model to predict traffic volumes in urban arterials. Subsequently, various variants of ARIMA have emerged [17–19]. Kalman filtering excels in regression problems. Guo et al. [20] applied an adaptive Kalman filtering model to predict short-term traffic flow. Hinsbergen et al. [21] used a localized extended Kalman filter (L-EKF) to estimate traffic states. In addition, traffic prediction methods based on Markov chains, exponential smoothing (ES), and Bayesian networks also perform well. For example, Qi et al. [22] proposed a hidden Markov model (HMM) to achieve short-term freeway traffic prediction during peak periods. Chan et al. [23] employed the hybrid exponential smoothing method and the Levenberg–Marquardt (LM) algorithm for short-term traffic flow forecasting. Wang et al. [24] used an improved Bayesian combination method (BCM) for short-term traffic flow prediction.

Statistical methods have some disadvantages, such as the inability to deal with nonlinear relationships between data. Machine learning methods, on the other hand, are more flexible. Machine learning methods are mainly divided into classical machine learning and deep learning.

Commonly used classical machine learning approaches include k-nearest neighbors (KNN), support-vector machine (SVM), random forest (RF), and decision tree (DT) methods. Cai et al. [25] proposed an improved KNN model to achieve short-term traffic multistep forecasting. Xu et al. [26] used kernel k-nearest neighbors (kernel-KNN) to predict road traffic states in time series. Cong et al. [27] presented a traffic flow prediction model based on the least squares support-vector machine, and automatically determined the least squares support-vector machine model with two parameters at the appropriate value by FOA. Xu et al. [28] used genetic programming (GP) and random forest (RF) techniques to achieve real-time crash prediction on freeways. Crosby et al. [29] proposed a spatially intensive decision tree for the prediction of traffic flow across the entire UK road network. Although classical machine learning methods are effective in identifying nonlinear relationships in traffic states, they still have many drawbacks, e.g., KNN models have low prediction accuracy for rare

categories and require high computational complexity when there are many features. It is difficult to choose a suitable kernel function by applying the SVM model. The random forests do not perform very well on high-dimensional sparse data. In addition, decision trees are prone to overfitting.

In order to solve the above problems, deep learning has been developed rapidly in recent years. The key to traffic prediction is to learn the temporal dependence and spatial dependence, where the methods to learn the temporal dependence are mainly recurrent neural networks (RNNs) and their variants long short-term memory (LSTM) and gated recurrent units (GRUs). Nejadettehad et al. [30] used three kinds of recurrent neural networks to predict short-term traffic flow. Van et al. [31] used recurrent neural networks to predict freeway travel time. Tian et al. [32] took advantage of LSTM to dynamically determine the optimal time lags to predict short-term traffic flow. Fu et al. [33] used LSTM and GRU methods to predict short-term traffic flow. These models consider the temporal dependence but ignore the spatial dependence in the road network. Therefore, they cannot accurately predict changes in the traffic state. Obtaining the temporal and spatial dependence is a prerequisite for accurate traffic prediction. There are also many models for the learning of spatial features. For example, Lv et al. [34] proposed a stacked autoencoder model to inherently learn the spatial and temporal correlations for traffic flow prediction. Yuan et al. [35] proposed a novel variable-wise weighted stacked autoencoder (VW-SAE) for hierarchical, layer-by-layer output-related feature representation. Ma et al. [36] proposed a convolutional neural network (CNN)-based model to learn traffic as images and predict large-scale, network-wide traffic speed. Wu et al. [37] proposed a model called CLTFP, which combines CNN and LSTM, to forecast future traffic flow. Jo et al. [38] adopted a convolutional neural network (CNN) to deal with map images representing traffic states and the model adopts images for both the input and the output of a CNN model to predict traffic speeds.

Although the above methods can handle spatial dependencies in traffic, CNNs are more suitable for Euclidean spatial structures such as pictures, and grids. Meanwhile, traffic road networks are complex networks, and the neighboring nodes are not fixed. Thus, the spatial features of the road network cannot be fully obtained by CNNs. In recent years, graph-based convolution operations have developed rapidly [39], and have become suitable for learning the structural features of graph types. He et al. [40] used LDA and GCN to tackle road link speed prediction. Li et al. [41] proposed a DCRNN model for obtaining spatio-temporal dependence in traffic flow forecasting; the model uses diffusion convolution to learn spatial dependence and a GRU to learn temporal dependence. Wu et al. [42] learned an adaptive dependency matrix via node embedding to obtain spatial dependency and temporal dependency through stacked dilated 1D convolution. Huang et al. [43] proposed a new graph attention network, cosAtt, to obtain spatial features through cosAtt and GCN and temporal features through a GLU. Roy et al. [44] consider important daily patterns and present-day patterns from traffic data in addition to spatio-temporal characteristics to improve the accuracy of predictions. However, these methods only consider the spatial features based on structure-aware graph embedding information, without considering the location information, so they cannot effectively obtain the spatial features.

2.2. Attention Mechanism. The attention mechanism has been a hot topic of neural network research in recent years, and it has been remarkable in neural machine translation, image captioning, time series prediction etc. The attention mechanism originates from the study of human vision, which determines which part of the input needs to be attended to and allocates processing resources to the important parts. Bahdanau et al. [45] proposed the use of an attention mechanism in the decoder to decide which part of the input sentence should be attended to. Xu et al. [46] introduced the application of soft and hard attention mechanisms to image captioning. Li et al. [47] proposed convolutional self-attention further improves Transformer' performance to achieve time series forecasting. Daiya et al. [48] proposed a multimodal deep learning architecture for stock movement prediction. Zhou et al. [49] used ProbSparse self-attention mechanism and distilling operation to handle quadratic time complexity and memory usage. In the area of traffic state prediction, prediction methods based on attention mechanisms are also developing rapidly. Park et al. [50] proposed the use of temporal attention, spatial attention and spatial sentinel vectors to obtain temporal and spatial dependencies. Wang et al. [51] proposed a novel spatial temporal graph neural network model for traffic flow prediction, and a learnable positional attention mechanism is applied in the model to aggregate information from adjacent roads. Guo et al. [52] proposed a novel attention-based spatio-temporal graph convolutional network (ASTGCN) to model recent, daily, and weekly dependencies.

Inspired by the above study, considering traffic location information and spatio-temporal characteristics, we learned both location- and structure-based information to obtain localized spatial features, learned localized temporal features through a GRU and, finally, considered the global spatio-temporal features of traffic networks through the attention mechanism.

3. Methodology

3.1. Data Processing. Given a speed sequence of data $T_0, T_1, T_2, \cdots, T_n$ with a length of n, the time interval is 5 minutes. To predict the future 15 minutes of data, for example, the input sample construction process of the model is shown in Figure 2. The input data of sample 1 is $\{T_0, T_1, T_2, \cdots, T_{11}\}$, and the label data is $\{T_{12}, T_{13}, T_{14}\}$. The input data of sample 2 is $\{T_1, T_2, T_3, \cdots, T_{12}\}$, and the label data is $\{T_{13}, T_{14}, T_{15}\}$. And so on, to obtain the entire input sample matrix. If predicting the next 30 minutes of data, the method is similar, i.e., the input data of sample 1 is unchanged, the label data is $\{T_{12}, T_{13}, T_{14}, T_{15}, T_{16}, T_{17}\}$, and the sample matrix is obtained recursively. The longer the prediction time, the more the label is increase.

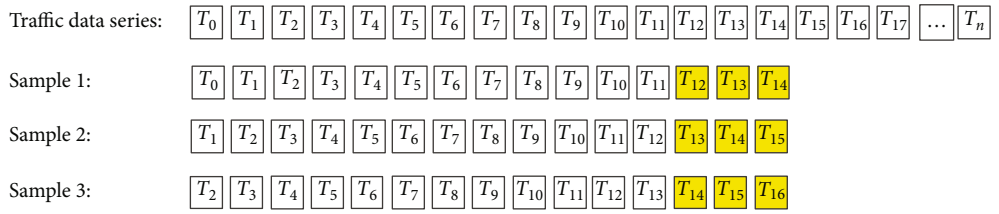


Figure 2: Sample construction.

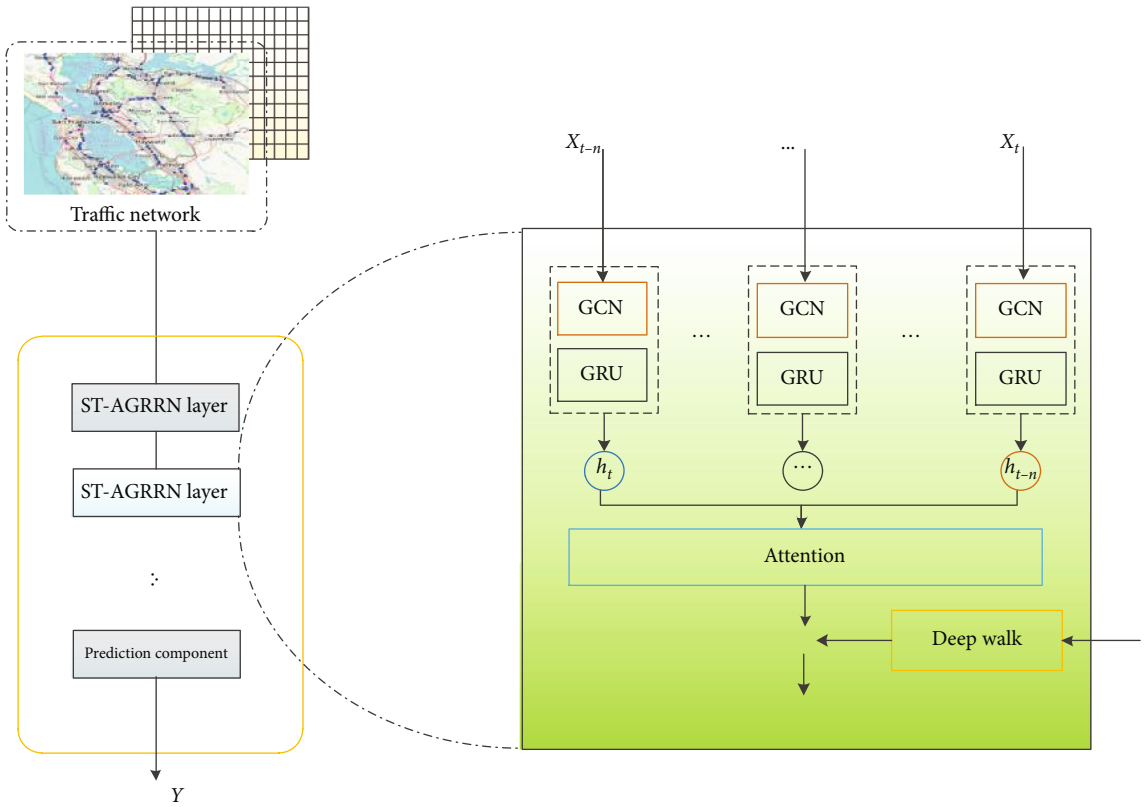


Figure 3: The architecture of the ST-AGRNN.

3.2. Traffic State Prediction Based on ST-AGRNN. The structure of the ST-AGRNN model is shown in Figure 3. In order to fully capture the localized spatial dependencies, we propose a new spatial feature extraction method by combining DeepWalk with a GCN, where DeepWalk obtains position-aware information and the GCN obtains structure-aware graph embedding information. The localized temporal dependencies are captured using the gated recurrent unit network, and the road network global spatio-temporal dependencies are captured using the attention mechanism. The specific details of each part of the model are presented in the next subsections.

3.2.1. Localized Spatial Dependency. Consider the urban road network as an undirected graph $G = (V, E)$, where V is the set of vertices in the graph and E is the set of edges. Denote the adjacency matrix of the graph by W. $D = \text{diag}(d_1, \cdots, d_n)$ denotes the degree matrix of the graph, where $d_i = \sum_{j=1}^{N} W_{ij}$ denotes the number of adjacencies of each vertex. Moreover, the Laplace matrix of the graph is expressed as $L = I_N - D^{-1/2}AD^{-1/2} = U\Lambda U^T$ (where U is an orthogonal matrix composed of eigenvectors), and the Fourier transform and inverse transform of the graph can be expressed as $\hat{x} = U^T x$ and $x = U\hat{x}$, respectively. A two-layer graph convolutional neural network can be represented as follows:

$$Z = f(X, A) = soft\max\left(\hat{A}\,\text{Re}\,LU\left(\hat{A}XW^{(0)}\right)W^{(1)}\right), \quad (1)$$

where X denotes the feature of the node, while A denotes the adjacency matrix of the graph. Calculated in the preprocessing step $\hat{A} = \tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$, where $\tilde{A} = A + I_N$ denotes the adjacency matrix with self-connections, $\widetilde{D_{ii}} = \sum_j A_{ij}$, $W^{(0)}$ is the weight of the input layer to the hidden layer, while $W^{(1)}$ is the weight of the hidden layer to the output layer.

The GCN aggregates information about neighboring nodes via convolution, which is a structure-based graph

Input: The training epoch N; the historical traffic state x_t; the traffic graph $G = (V, E)$; the window size of historical traffic state p; the predicted length of traffic state q;
Output: Learned ST-AGRNN model
1: Initialization parameter θ;
2: Data processing;
3: For $\forall i \in N$ do
4: Select real historical data $x_{t-p}, \cdots, x_t$;
5: Select real future data $x_{t+1}, \cdots, x_{t+q}$;
6: Input real historical data $x_{t-p}, \cdots, x_t$ and the traffic graph $G = (V, E)$ into GCN and GRU to get h_i;
7: Input h_i into attention to get c_i;
8: Use DeepWalk on G and get the embedding result $\tilde{s}$;
9: Concatenate c_i and $\tilde{s}$, $o_t^i = c_i \oplus \tilde{s}$;
10: Optimize θ by minimizing the loss function;
11: End for

Algorithm 1: Training of ST-AGRNN.

Table 1: Traffic speed prediction performance under different benchmark methods in the PeMSD4 and PeMSD8 datasets (bold is the best; underline is the second best.).

Model	PeMSD4 (MAE/RMSE/MAPE(%))		
	15 min	30 min	60 min
HA	2.54/4.96/5.56	2.54/4.96/5.56	2.54/4.96/5.56
ARIMA(2003)	2.51/5.72/5.32	2.75/6.34/5.69	3.21/7.36/6.56
DCRNN(2018)	1.35/2.94/2.68	1.77/4.06/3.71	2.26/5.28/5.10
STGCN(2018)	1.47/3.01/2.92	1.93/4.21/3.98	2.55/5.65/5.39
ASTGCN(2019)	2.12/3.96/4.16	2.42/4.59/4.80	2.73/5.21/5.46
GWN(2019)	1.30/2.68/2.67	1.70/3.82/3.73	2.03/4.65/4.60
LSGCN(2020)	1.45/2.93/2.90	1.82/3.92/3.84	2.22/4.83/4.85
USTGCN(2021)	1.40/2.69/2.81/	1.64/3.19/3.23	2.03/4.25/4.32
ST-AGRNN	**1.19/2.36/2.17**	**1.45/2.98/2.69**	**1.76/3.63/3.24**
Model	PeMSD8 (MAE/RMSE/MAPE(%))		
	15 min	30 min	60 min
HA	1.98/4.11/3.94	1.98/4.11/3.94	1.98/4.11/3.94
ARIMA(2003)	1.90/4.87/5.11	2.12/5.24/5.21	2.79/6.22/5.62
DCRNN(2018)	1.17/2.59/2.32	1.49/3.56/3.21	1.87/4.50/4.28
STGCN(2018)	1.19/2.62/2.34	1.59/3.61/3.24	2.25/4.68/4.54
ASTGCN(2019)	1.49/3.18/3.16	1.67/3.69/3.59	1.89/4.13/4.22
LSGCN(2020)	1.16/2.45/2.24	1.46/3.28/3.02	1.81/4.11/3.89
USTGCN(2021)	1.14/2.15/2.07	1.25/**2.58**/2.35	1.70/**3.27**/3.22
ST-AGRNN	**1.015/2.07/1.82**	**1.24**/2.63/**2.21**	**1.53**/3.33/**2.71**

embedding algorithm. The obtained embedding representation cannot retain the position relationship between nodes, which is a very important relationship between nodes in the traffic network. Deepwalk's objective function forces nodes that are close in the shortest path to be close in the embedding space representation [53]. In order to fully exploit the spatial features of the road network, we introduce the DeepWalk algorithm to learn the position embedding representation between nodes.

The graph embedding algorithm based on the random walk is also close in the embedding space for nodes that are close in the shortest path. This allows the resulting embedding space to also preserve the relative positional relationships. These relations are an important complement to the structure-based embedding space, and are necessary for spatial features in traffic.

The random walk with v_t as the vertex is represented as $\{W_{v_t}^1, W_{v_t}^2, \cdots, W_{v_t}^k\}$, where $W_{v_t}^k$ denotes the k th node in the path with v_t as the root. For all of the nodes in the graph, each node has another similar path. We then obtain a sequence matrix W. The corresponding graph embedding representation containing the location information is then obtained by the

Table 2: Traffic flow prediction performance under different benchmark methods in the PeMSD4 and PeMSD8 datasets (bold is the best; underline is the second best.).

Model	PeMSD4		
	MAE	RMSE	MAPE (%)
HA	38.03	59.24	27.88
ARIMA(2003)	33.73	48.80	24.18
STGCN(2018)	21.16	34.89	13.83
DCRNN(2018)	21.22	33.44	14.17
ASTGCN(r)(2019)	22.93	35.22	16.56
GWN(2019)	24.89	39.66	17.29
LSGCN(2020)	21.53	33.86	13.18
STSGCN(2020)	21.19	33.65	13.90
STFGNN(2021)	20.48	32.51	16.77
Z-GCNETs(2021)	19.50	31.61	12.78
STG-NCDE(2022)	19.21	31.09	**12.76**
ST-AGRNN(ours)	**18.97**	**30.003**	12.81
Model	PeMSD8		
	MAE	RMSE	MAPE (%)
HA	34.86	59.24	27.88
ARIMA(2003)	31.09	44.32	22.73
STGCN(2018)	17.50	27.09	11.29
DCRNN(2018)	16.82	26.36	10.92
ASTGCN(r) (2019)	18.25	28.06	11.64
GWN(2019)	18.28	30.05	12.15
LSGCN(2020)	17.73	26.76	11.20
STSGCN(2020)	17.13	26.80	10.96
STFGNN(2021)	16.94	26.25	10.60
Z-GCNETs(2021)	15.75	25.11	10.01
STG-NCDE(2022)	15.45	24.81	9.92
ST-AGRNN(ours)	**14.95**	**23.15**	**9.21**

update procedure—the skip-gram algorithm. The embedding representation is denoted as s, and then the final result is obtained by the fully connected layer.

$$\tilde{s} = f(W \cdot s + b), \tag{2}$$

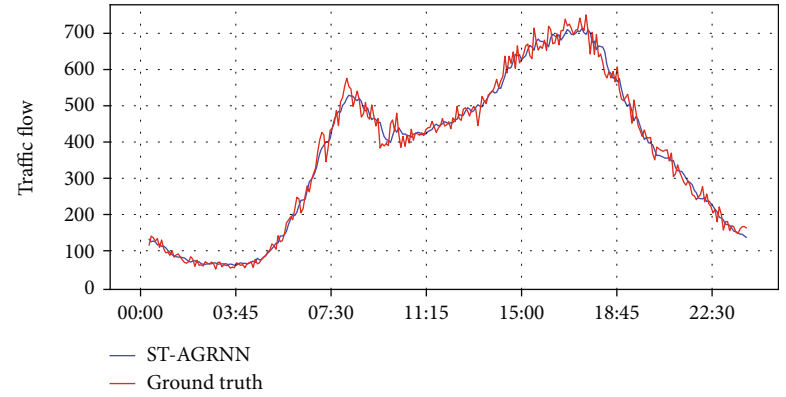
where $\tilde{s}$ denotes the graph embedding representation, while W and b are the learnable weights and biases, respectively.

3.2.2. Localized Temporal Feature. Temporal dependence is another major problem in traffic prediction. Recurrent neural network (RNN) models are very effective for time-series data processing, but they suffer from gradient disappearance and gradient explosion. GRUs and LSTM are variants of RNN that can effectively overcome these problems.

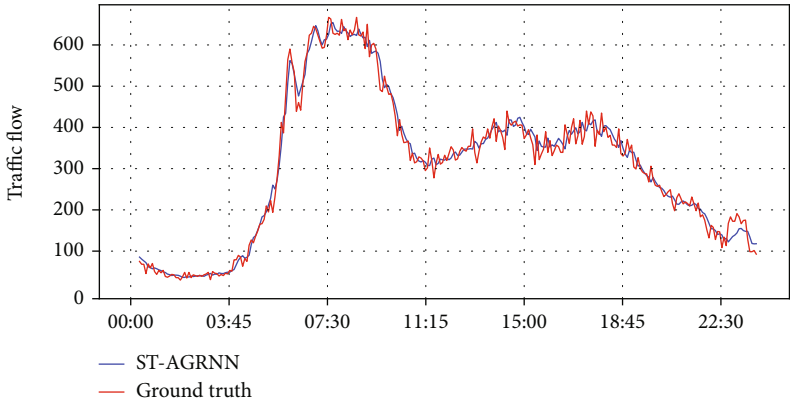
GRU is used to handle temporal dependence. s_t is the output of GCN at time t, x_t is the traffic state at the present moment, and r_t is the reset gate that determines whether the previous moment information is retained or not—if it is 1, then the message is carried to the next moment; if it is 0, then the message is ignored. h_{t-1} is the hidden state at the previous moment. z_t is the update gate, which is a value between 0 and 1 that determines how much information is remembered from the previous moment—if it is 1, then more information is remembered; if it is 0, then more is forgotten. $\widetilde{h}_t$ is the current memory content, and h_t is the output of the current moment.

$$\begin{aligned} s_t &= \mathrm{GCN}(x_t), \\ r_t &= \sigma(W_r \cdot [h_{t-1}, s_t \cdot x_t] + b_r), \\ z_t &= \sigma(W_z \cdot [h_{t-1}, s_t \cdot x_t] + b_z), \\ \widetilde{h}_t &= \tanh\left(W_h \cdot [r_t \odot h_{t-1}, s_t \cdot x_t] + b_h\right), \\ h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \widetilde{h}_t. \end{aligned} \tag{3}$$

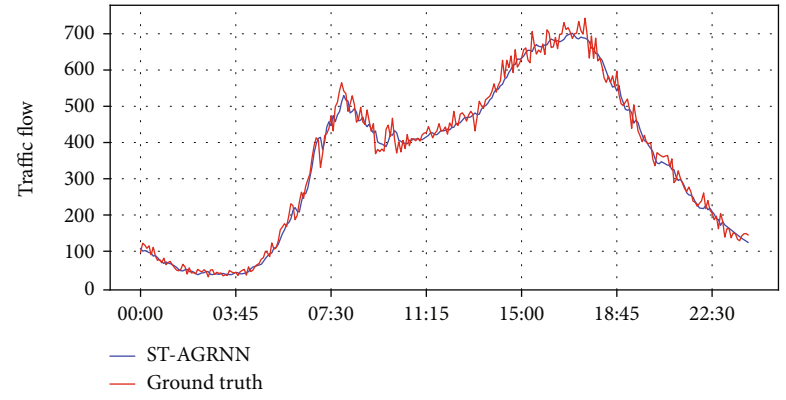
3.2.3. Global Spatio-Temporal Correlations. Critical intersections in cities often have a large impact on regional traffic, and congestion at critical intersections is likely to evolve into congestion in the associated areas. In order to strengthen the modeling ability of traffic networks, this paper obtains global spatio-temporal correlations through the attention mechanism. All of the hidden states of the GRU network are used as the input of the attention network, and then the weights of each hidden state of the GRU are calculated to obtain



(a) Node 111 in PeMSD4 for 15 min

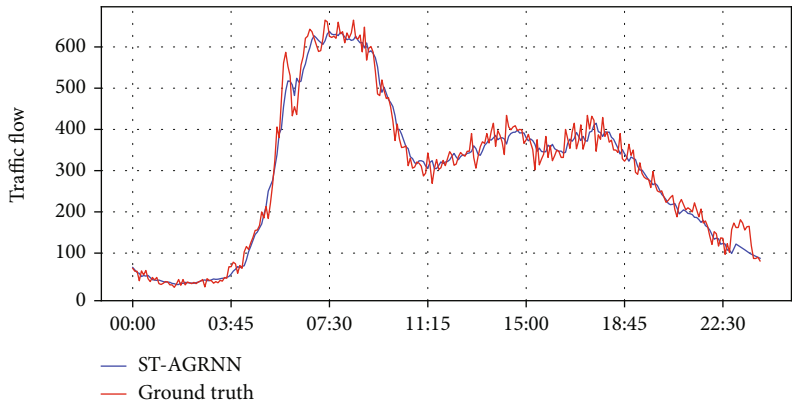


(b) Node 261 in PeMSD4 for 15 min



(c) Node 111 in PeMSD4 for 60 min

Figure 4: Continued.



(d) Node 261 in PeMSD4 for 60 min

FIGURE 4: Traffic flow forecast visualization in PeMSD4.

the traffic information changes in the road network at each moment. The attention network is calculated as follows:

$$e_i = W^{(1)}\left(\tanh\left(W^{(0)}h_i + b^{(0)}\right)\right) + b^{(1)},$$
$$a_i = \frac{\exp(e_i)}{\sum_{k=1}^{n}\exp(e_k)}, \tag{4}$$
$$c_i = \sum_{i=1}^{n} a_i * h_i,$$

where e_i is the attention coefficient, h_i is the GRU hidden state, $W^{(0)}$ and $W^{(1)}$ are the trainable weight parameters, $b^{(0)}$ and $b^{(1)}$ are the trainable bias values, a_i is the normalized attention coefficient, and c_i is the attention weight.

3.3. Prediction Component. We predict future changes in traffic state based on historical traffic states. In the prediction component, we concatenate the attention mechanism and the location-based graph embedding output as follows:

$$o_t^i = c_i \oplus \tilde{s}. \tag{5}$$

The concatenation result is used as the input of the fully connected layer, and the final traffic state is obtained by the sigmoid activation function. It is expressed as y_{t+T}^i, where T is the predicted time step, in the following form:

$$y_{t+T}^i = \text{sigmoid}\left(W_s o_t^i + b_s\right), \tag{6}$$

where W_s and b_s are the learnable weights and biases, respectively.

The training overview of the model is shown in Algorithm 1. We used Adam to optimize the model. We used TensorFlow to implement the proposed model.

4. Experiments

4.1. Experimental Settings. The software and hardware environments for the experiments were configured as follows: PYTHON 3.6.2, NUMPY 1.16.0, TENSORFLOW 1.14.0, and Memory: 64 GB.

For this paper, we used speed and traffic flow to represent traffic states, where 80% of the data were used as the training set and 20% as the test set. In the experiments, the speed was predicted for 15, 30, and 60 minutes, and the flow prediction was predicted from 5 to 60 minutes with 12 time windows.

We use the root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage errors (MAPE) to evaluate the models.

4.2. Dataset Description. In the experiment, we used two real-world traffic datasets: PeMSD4, and PeMSD8 [43].

PeMSD4 was collected from the Caltrans Performance Measurement System (PeMS) and the traffic data in the San Francisco Bay Area, with 307 sensors on 29 roads. The dataset spanned from January to February 2018.

PeMSD8 refers to the traffic data in San Bernardino from July to August 2016, with 170 detectors on 8 roads.

4.3. Baselines. In this paper, the traffic state includes traffic speed and flow. For the traffic speed, we used the proposed model to predict 15, 30, and 60 minutes. The compared baseline models contain both traditional HA and ARIMA, along with neural network models such as STGCN [7], DCRNN [41], ASTGCN [52], GWN [42], LSGCN [43], and USTGCN [44].

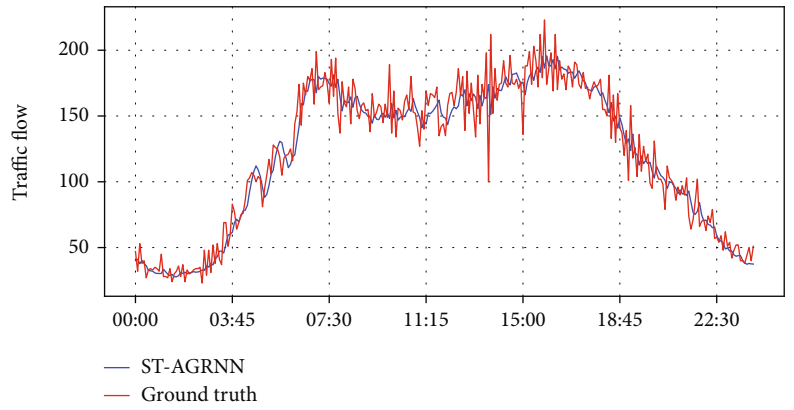
In traffic flow forecasting, all models have a prediction window from 1 to 12, i.e., a prediction time from 5 minutes to 60 minutes, in 5-minute intervals. The baseline models compared included both traditional and neural network models, for a total of 11.

The details of the baseline model are as follows:

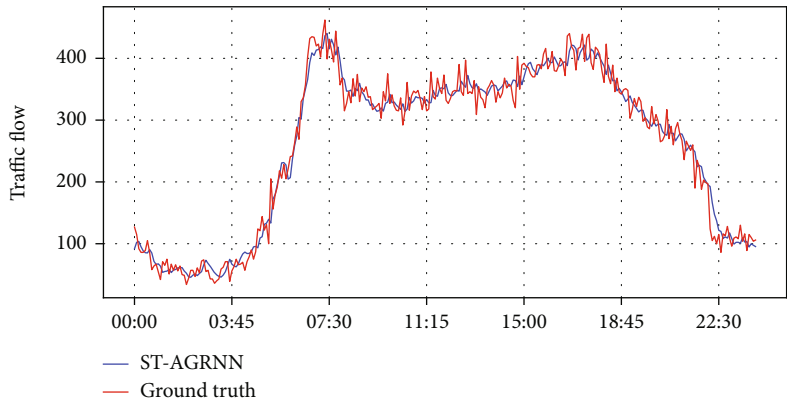
(1) HA: the average traffic information of the previous period is used as the forecast value

(2) ARIMA: autoregressive integrated moving average

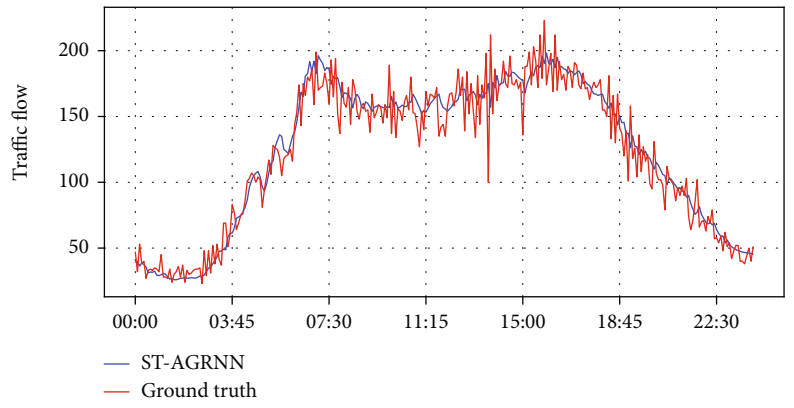
(3) STGCN: spatio-temporal graph convolutional network, which consists of several spatio-temporal convolutional blocks



(a) Node 9 in PeMSD8 for 15 min

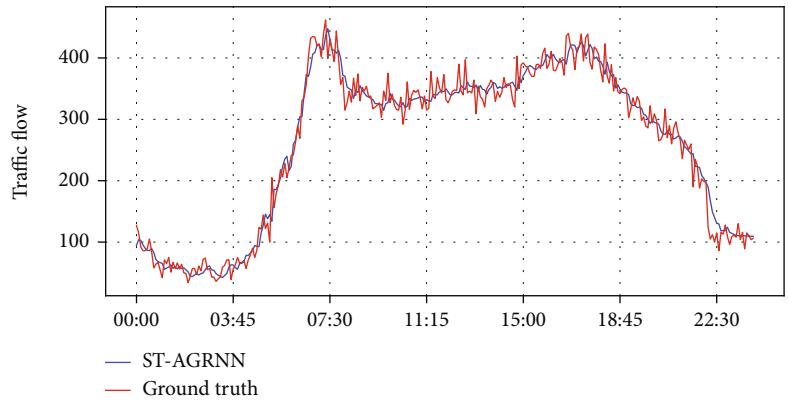


(b) Node 112 in PeMSD8 for 15 min



(c) Node 9 in PeMSD8 for 60 min

Figure 5: Continued.



(d) Node 112 in PeMSD8 for 60 min

Figure 5: Traffic flow forecast visualization in PeMSD8.

(4) DCRNN: diffusion convolutional recurrent neural network, which obtains spatial dependencies through bidirectional random walks and temporal dependencies through an encoder–decoder structure with scheduled sampling

(5) ASTGCN(r): three independent components with the same structure are used to obtain the recent, daily, and weekly dependencies in the traffic data. The spatio-temporal attention mechanism and spatio-temporal convolution are used to obtain the spatio-temporal dependencies within the components. For the sake of experimental fairness, only the recent components are used

(6) GWN: a new adaptive dependency matrix is learned by node embedding to capture the hidden spatial dependencies in the data and obtain temporal dependence via a stacked dilated 1D convolutional component

(7) LSGCN: the model uses spatial gated block and gated linear units (GLU) convolution to capture spatio-temporal features

(8) USTGCN: the model obtains complex spatio-temporal correlations through the proposed unified spatio-temporal convolution strategy

(9) STSGCN [54]: spatio-temporal synchronous graph convolutional network, which uses a spatio-temporal synchronous graph convolutional module to capture the complex localized spatio-temporal correlations and deploys multiple modules to capture the heterogeneities in localized spatio-temporal network series

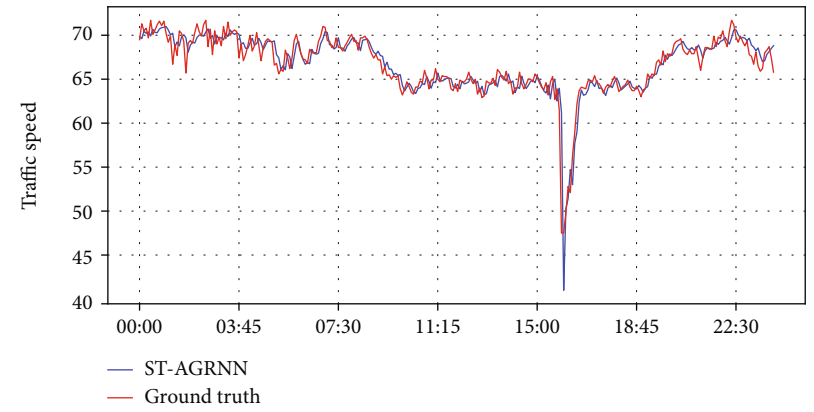
(10) STFGNN [55]: spatio-temporal fusion graph neural network, which uses spatio-temporal fusion graph neural modules and a gated CNN module to capture the spatio-temporal correlations

(11) Z-GCNETs [56]: Z-GCNETs introduce new GCNs with a time-aware zigzag topological layer

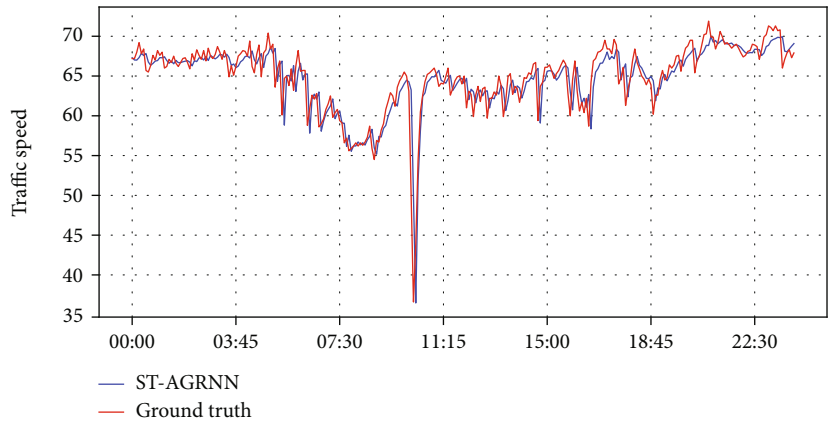
(12) STG-NCDE [57]: spatio-temporal graph neural controlled differential equation, which extends the concept and designs two NCDEs to capture the spatio-temporal correlations

4.4. Experimental Results. The traffic state prediction results for all baseline models and our model are shown in Tables 1 and 2. In Table 1, we can see that our proposed model performs better overall on the datasets PeMSD4 and PeMSD8 compared to the other baseline models for 15-, 30-, and 60-minute traffic speed predictions. Taking the 15-minute speed forecast as an example, on the PeMSD4 dataset, our model is better than HA, ARIMA, DCRNN, STGCN, ASTGCN, GWN, LSGCN, and USTGCN with 53.14, 52.58, 11.85, 19.04, 43.86, 8.46, 17.93, and 15% lower MAE, with 52.41, 58.74,19.72, 21.59, 40.40, 11.94, 19.45, and 12.26% lower RMSE, and with 60.97, 59.21, 19.02, 25.68, 47.83, 18.72, 25.17, and 22.77% lower MAPE, respectively. On the PeMSD8 dataset, our model is better than HA, ARIMA, DCRNN, STGCN, ASTGCN, LSGCN, and USTGCN with 48.73, 46.57, 13.24, 14.7, 31.87, 12.5, and 10.96% lower MAE, with 49.63, 57.49, 20.07, 20.99, 34.9, 15.51, and 3.72% lower RMSE, and with 53.8, 64.38, 21.55, 22.22, 42.4, 18.75, and 12.07% lower MAPE, respectively. From the results, it is clear that ST-AGRNN performs well in both short- and long-term predictions. In particular, on the PeMSD4 dataset, the ST-AGRNN model is optimal on all three-evaluation metrics. Except for the RMSE metric, which is the second best on the PeMSD8 dataset, the other metrics are also optimal for long- and short-term prediction.

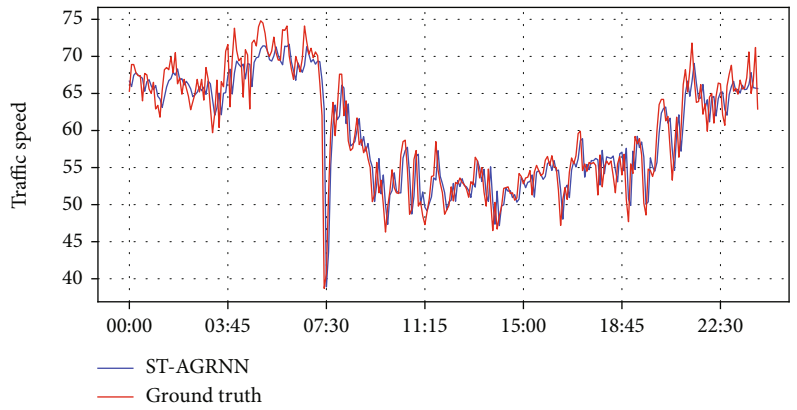
HA and ARIMA are the worst performers because they do not capture spatio-temporal correlations effectively. Since STGCN has cumulative errors, it does not perform as well as DCRNN. DCRNN can effectively obtain complex spatial correlations through diffusion convolution. ASTGCN considers the periodicity of prediction, so it is better than STGCN for long-term prediction.



(a) Node 111 in PeMSD4 for 15 min

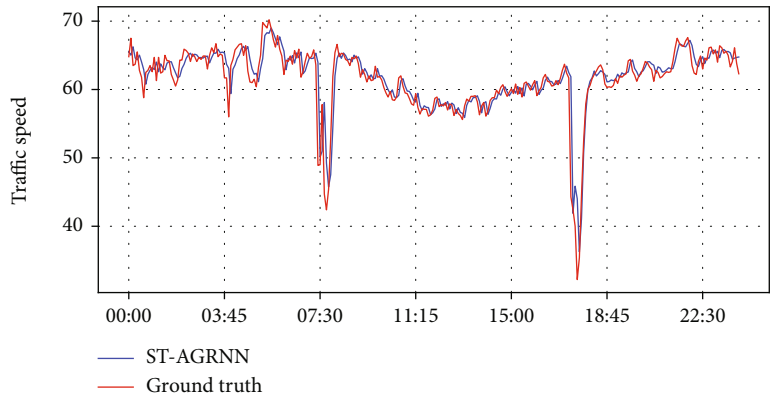


(b) Node 261 in PeMSD4 for 15 min



(c) Node 9 in PeMSD8 for 15 min

Figure 6: Continued.



(d) Node 112 in PeMSD8 for 15 min

Figure 6: Speed forecast visualization in PeMSD4 and PeMSD8.

The spatial gate block of LSGCN integrates the proposed cosAtt and GCN, and in combination with a GLU can effectively extract complex spatio-temporal correlations. Meanwhile, the USTGCN model considers the important historical and present-day patterns in traffic data, in addition to the unified spatio-temporal convolution strategy. Therefore, its prediction performance is the second best.

Table 2 shows the results of traffic flow forecasting performed from 5 minutes all the way to 60 minutes, with a prediction window from 1 to 12, and all of the results are averaged. Compared with all of the baseline models, our proposed model performs the best in traffic flow prediction. From table 2, on the PeMSD4 dataset, our model is better than HA, ARIMA, STGCN, DCRNN, ASTGCN(r), GWN, LSGCN, STSGCN, STFGNN, Z-GCNETs, and STG-NCDE with 50.11, 43.75, 10.34, 10.60, 17.26, 23.78, 11.89, 10.47, 7.37, 2.71, and 1.24% lower MAE, with 49.35, 38.51, 14, 10.27, 14.81, 24.34, 11.39, 10.83, 7.71, 5.08, and 3.49% lower RMSE, and with 54.05, 47.02, 7.37, 9.59, 22.64, 25.91, 2.8, 7.84, 23.61, -0.23, and -0.39% lower MAPE, respectively. On the PeMSD8 dataset, our model is better with 57.11, 51.91, 14.57, 11.11, 18.08, 18.21, 15.67, 12.72, 11.74, 5.07, and 3.23% lower MAE, with 60.92, 47.76, 14.54, 12.17, 17.49, 22.96, 13.49, 13.61, 11.8, 7.8, and 6.69% lower RMSE, and with 66.96, 59.48, 18.42, 15.65, 20.87, 24.19, 17.76, 15.96, 13.11, 7.99, and 7.15% lower MAPE, respectively.

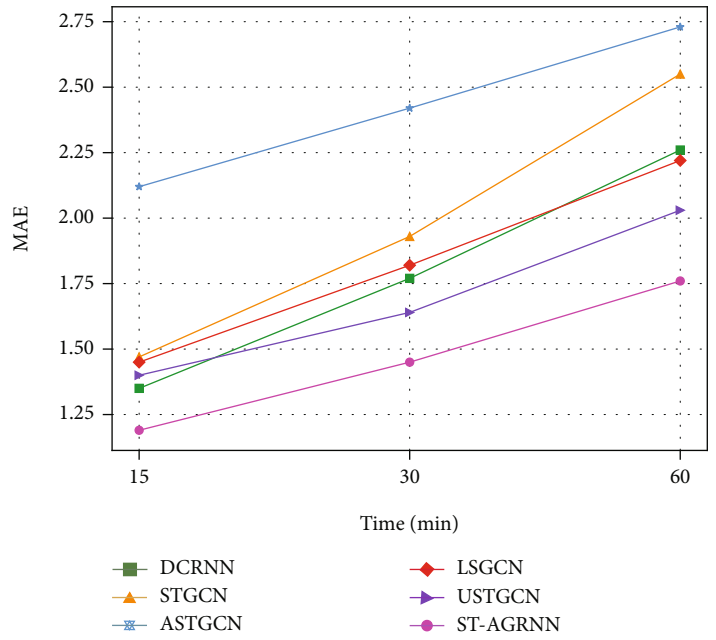
The STSGCN model considers both localized spatio-temporal correlations and the heterogeneities in spatial-temporal data. Therefore, its performance is better than STGCN, DCRNN, ASTGCN(R), GWN, and LSGCN. The SFTGNN obtains hidden spatio-temporal correlations by fusing spatial and temporal graph operations and integrating the gate convolution module at the same time. Z-GCNETs proposed new GCNs with a time-aware Zigzag topological layer to obtain spatio-temporal correlation. The STG-NCDE model uses two neural controlled differential equations (NCDEs) to obtain the temporal and spatial correlations. Since The STSGCN model only extracted localized spatio-temporal correlations, its performance was inferior to that of SFTGNN, Z-GCNets, and STG-NCDE. The ST-AGRNN model obtains both localized and global spatio-temporal correlation and combines location-based graph embedding representation to obtain localized spatial correlation. So, the overall performance on both datasets is better than all baseline models.

4.5. Case Study. We selected two nodes with heavy traffic from the two datasets to show the ground-truth and predicted curves: nodes 111 and 261 in PeMDS4 and nodes 9 and 112 in PeMSD8, as shown in Figures 4 and 5, respectively. From the figures, it can be seen that the model fits this trend well in places with huge traffic flows between 7 : 00 and 9 : 00 a.m. and between 3 : 00 and 6 : 40 p.m. Figure 6 shows the change in the nodes' 15-minute speed. From the figure, the traffic speed also drops sharply at the peak time of corresponding traffic flow.

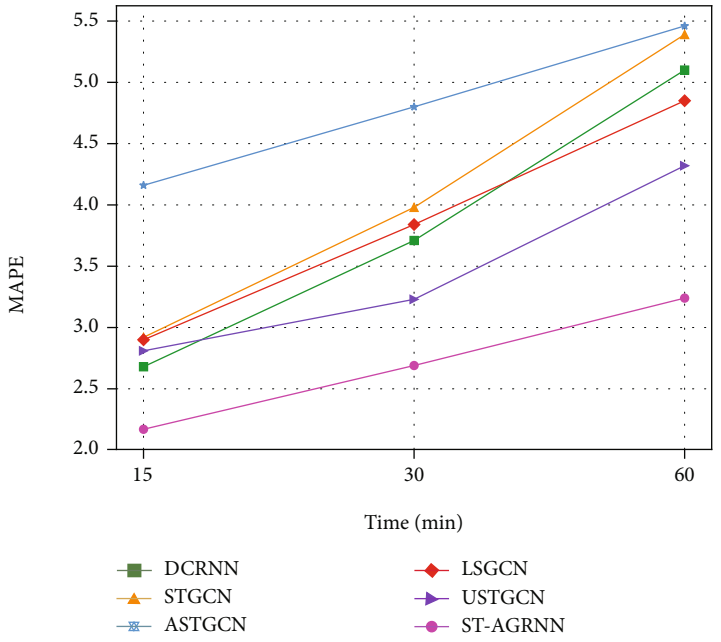
4.6. Error for each Length of Forecasting. Figure 7 shows the trend of the prediction error of the model in terms of prediction speed on two datasets. From the figure, it can be seen that although the error increases for all of the models as the prediction length increases, the error of our model is smaller than baselines and the increasing trend of our model is the flattest. This proves that our model is more stable than the baseline models.

4.7. Ablation Experiments. In the traffic network, the road sections at different locations play different roles in traffic. Road sections in central areas have a greater impact on the surrounding traffic, while remote road sections play a small role in influencing traffic. These are the global spatio-temporal correlations. To verify the importance of global spatio-temporal correlations, we conduct ablation experiments on speed prediction.

From the comparison of the traffic speed prediction results in Table 3, it can be seen that the prediction error of the ST-AGRNN model with the attention mechanism is smaller overall than the error of ST-DWGRU [58] without the attention mechanism. As an example of the 60-minute prediction results, the MAE of ST-AGRNN on the PeMSD4 dataset is 7.3% smaller than that of ST-DWGRU, the RMSE is 9.4% smaller, and the MAPE is 8.2% smaller. The MAE of ST-AGRNN on the PeMSD8 dataset is 2.5% smaller than that of ST-DWGRU, the RMSE is 4.5% smaller, and the

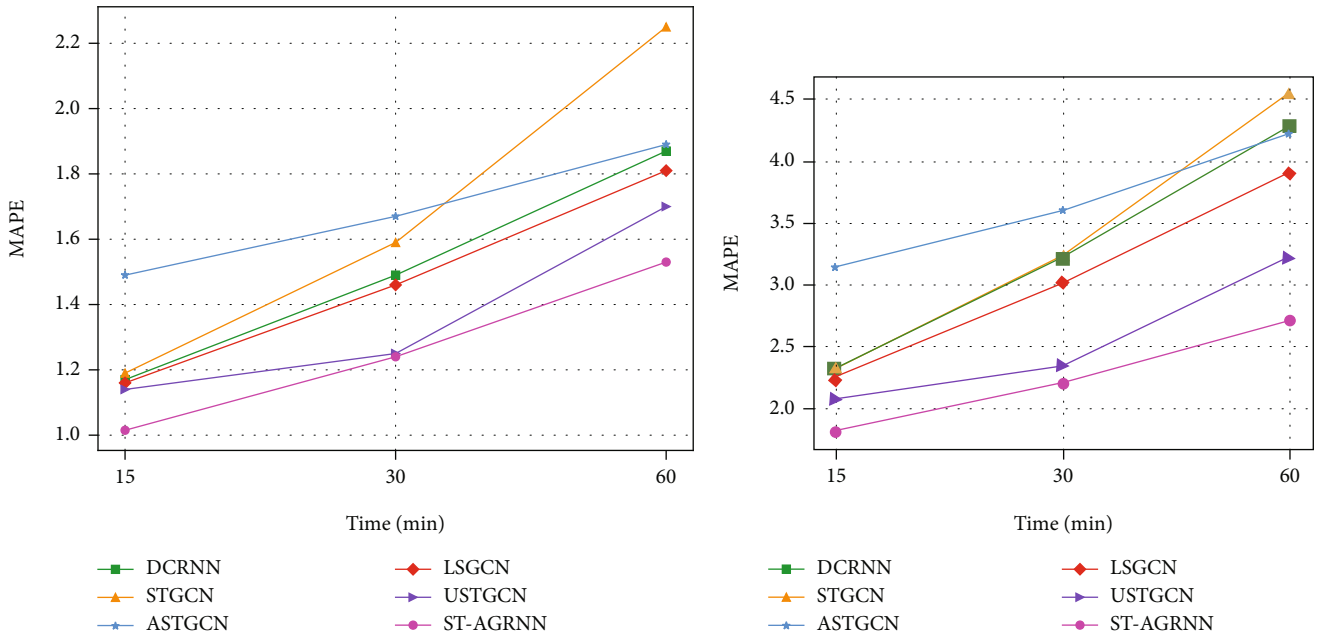


(a) MAE for speed forecast on PeMSD4



(b) MAPE for speed forecast on PeMSD4

Figure 7: Continued.



(c) MAE for speed forecast on PeMSD8

(d) MAPE for speed forecast on PeMSD8

Figure 7: Prediction error trend.

Table 3: Comparison of traffic speed prediction results of the ST-AGRRG and ST-DWGRU models (bold is the best).

T	Metric	PeMSD4		PeMSD8	
		ST-AGRNN	ST-DWGRU	ST-AGRNN	ST-DWGRU
15 min	MAE	**1.19**	1.20	1.015	**1.005**
	RMSE	**2.36**	2.40	**2.07**	2.08
	MAPE	**2.17**	2.21	1.82	**1.81**
30 min	MAE	**1.45**	1.48	**1.24**	1.25
	RMSE	**2.98**	3.12	**2.63**	2.70
	MAPE	**2.69**	2.75	**2.21**	2.24
60 min	MAE	**1.76**	1.90	**1.53**	1.57
	RMSE	**3.63**	4.01	**3.33**	3.49
	MAPE	**3.24**	3.53	**2.71**	2.78

MAPE is 2.5% smaller. From the results, it is clear that the ST-AGRNN model is more effective in obtaining complex spatio-temporal information.

5. Conclusions

A new traffic state prediction model is proposed, in which localized spatial correlation is obtained by a GCN and Deep-Walk, localized temporal correlation is obtained by a GRU, and the global spatio-temporal correlations is obtained by the attention mechanism. Finally, the proposed model ST-AGRNN was tested with two publicly available datasets, namely, PeMSD4 and PeMSD8. In terms of traffic speed prediction, MAE improved by 15-53.14% and 10.96-48.73%, RMSE improved by 12.26-52.41% and 3.72-49.63%, and MAPE improved by 22.77-60.97% and 12.07-53.8% on the PEMSD4 and PEMSD8 datasets, respectively, compared to the baseline models. Meanwhile, the ST-AGRNN model also showed different degrees of improvement in traffic flow prediction compared with the baseline models. From the results, it is clear that ST-AGRNN outperforms all of the baseline models, and is more stable.

Data Availability

Previously reported traffic data that were used to support the study are available. These prior studies (and datasets) are cited at relevant places within the text as references [43].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was financially supported by the National Natural Science Foundation of China (61977001).

References

[1] Y. Lv, Y. Duan, W. Kang, Z. Li, and F. Wang, "Traffic flow prediction with big data: a deep learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865–873, 2014.

[2] R. Yasdi, "Prediction of road traffic using a neural network approach," *Neural Computing & Applications*, vol. 8, no. 2, pp. 135–142, 1999.

[3] D. Xu, P. Peng, C. Wei, D. He, and Q. Xuan, "Road traffic network state prediction based on a generative adversarial network," *IET Intelligent Transport Systems*, vol. 14, no. 10, pp. 1286–1294, 2020.

[4] H. Xue and F. D. Salim, "TERMCast: Temporal Relation Modeling for Effective Urban Flow Forecasting," in *Advances in Knowledge Discovery and Data Mining. PAKDD 2021*, vol. 12712 of Lecture Notes in Computer Science, Springer, Cham.

[5] L. Cai, K. Janowicz, G. Mai, B. Yan, and R. Zhu, "Traffic transformer: capturing the continuity and periodicity of time series for traffic forecasting," *Transactions in GIS*, vol. 24, no. 3, pp. 736–755, 2020.

[6] X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang, and Y. Wang, "Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction," *Sensors*, vol. 17, no. 4, p. 818, 2017.

[7] Y. Bing, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional neural network: a deep learning framework for traffic forecasting," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3634–3640, Stockholm, Sweden, 2018.

[8] M.-T. Luong, H. Pham, and C. D. Manning, "Effective Approaches to Attention-Based Neural Machine Translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, Lisbon, Portugal, September 2015.

[9] B. Chen, W. Deng, and J. Hu, "Mixed highorder attention network for person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 371–381, Seoul, Korea (South), 2019.

[10] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li, "Single Shot Text Detector with Regional Attention," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3066–3074, Venice, Italy, October 2017.

[11] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 167–174, Scottsdale, AZ, USA, December 2015.

[12] J. Bai, J. Zhu, Y. Song et al., "A3T-GCN: attention temporal graph convolutional network for traffic forecasting," *ISPRS International Journal of Geo-Information*, vol. 10, no. 7, p. 485, 2021.

[13] I. Padhi, Y. Schiff, I. Melnyk et al., "Tabular Transformers for Modeling Multivariate Time Series," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3565–3569, Toronto, ON, Canada, June 2021.

[14] S. Wu, X. Xiao, Q. Ding, P. Zhao, Y. Wei, and J. Huang, "Adversarial sparse transformer for time series forecasting," in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*, pp. 17105–17115, Vancouver BC Canada, December 2020.

[15] M. S. Ahmed and A. R. Cook, "Analysis of freeway traffic time-series data by using box-Jenkins techniques," *Transportation Research Record*, vol. 722, pp. 1–9, 1979.

[16] M. M. Hamed, H. R. Al-Masaeid, and Z. M. B. Said, "Short-Term prediction of traffic volume in urban arterials," *Journal of Transportation Engineering*, vol. 121, no. 3, pp. 249–254, 1995.

[17] Q. Y. Ding, X. F. Wang, X. Y. Zhang, and Z. Q. Sun, "Forecasting traffic volume with space-time ARIMA model," *Advanced Materials Research*, vol. 156-157, pp. 979–983, 2010.

[18] M. Van Der Voort, M. Dougherty, and S. Watson, "Combining kohonen maps with arima time series models to forecast traffic flow," *Transportation Research Part C: Emerging Technologies*, vol. 4, no. 5, pp. 307–318, 1996.

[19] B. M. Williams and L. A. Hoel, "Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: theoretical basis and empirical results," *Journal of Transportation Engineering*, vol. 129, no. 6, pp. 664–672, 2003.

[20] J. Guo, W. Huang, and B. M. Williams, "Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification," *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 50–64, 2014.

[21] C. P. I. J. V. Hinsbergen, T. Schreiter, F. S. Zuurbier, J. W. C. V. Lint, and H. J. V. Zuylen, "Localized extended kalman filter for scalable real-time traffic state estimation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 1, pp. 385–394, 2012.

[22] Y. Qi and S. Ishak, "A hidden markov model for short term prediction of traffic conditions on freeways," *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 95–111, 2014.

[23] K. Y. Chan, T. S. Dillon, J. Singh, and E. Chang, "Neural-Network-Based Models for Short-Term Traffic Flow Forecasting Using a Hybrid Exponential Smoothing and Levenberg–Marquardt Algorithm," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 2, pp. 644–654, 2012.

[24] J. Wang, W. Deng, and Y. Guo, "New Bayesian combination method for short-term traffic flow forecasting," *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 79–94, 2014.

[25] P. Cai, Y. Wang, G. Lu, P. Chen, C. Ding, and J. Sun, "A spatiotemporal correlative _k_ -nearest neighbor model for short-term traffic multistep forecasting," *Transportation Research Part C: Emerging Technologies*, vol. 62, pp. 21–34, 2016.

[26] X. Dongwei, Y. Wang, P. Peng, S. Beilun, Z. Deng, and H. Guo, "Real-time road traffic state prediction based on kernel-KNN," *Transportmetrica A: Transport Science*, vol. 16, no. 1, pp. 104–118, 2020.

[27] Y. Cong, J. Wang, and X. Li, "Traffic flow forecasting by a least squares support vector machine with a fruit fly optimization algorithm," *Procedia Engineering*, vol. 137, pp. 59–68, 2016.

[28] C. Xu, W. Wang, and P. Liu, "A genetic programming model for real-time crash prediction on freeways," *IEEE Transactions*

on Intelligent Transportation Systems, vol. 14, no. 2, pp. 574–586, 2013.

[29] H. Crosby, S. A. Jarvis, and P. Davis, "Spatially-Intensive Decision Tree Prediction of Traffic Flow across the Entire UK Road Network," in *Proceedings of the 2016 IEEE/ACM 20th international symposium on distributed simulation and real time applications*, pp. 116–119, London, UK, September 2016.

[30] A. Nejadettehad, H. Mahini, and B. Bahrak, "Short-term demand forecasting for online car-hailing services using recurrent neural networks," *Applied Artificial Intelligence*, vol. 34, no. 9, pp. 674–689, 2020.

[31] J. W. Van Lint, S. P. Hoogendoorn, and H. J. van Zuylen, "Freeway travel time prediction with state-space neural networks: modeling state-space dynamics with recurrent neural networks," *Transportation Research Record*, vol. 1811, no. 1, pp. 30–39, 2002.

[32] Y. Tian and L. Pan, "Predicting Short-Term Traffic Flow by Long Short-Term Memory Recurrent Neural Network," in *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, pp. 153–158, Chengdu, China, December 2015.

[33] R. Fu, Z. Zhang, and L. Li, "Using LSTM and GRU Neural Network Methods for Traffic Flow Prediction," in *2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pp. 324–328, Wuhan, China, November 2016.

[34] J. Chen, W. Yuan, J. Cao, and H. Lv, "Traffic-flow prediction via granular computing and stacked autoencoder," *Granular Computing*, vol. 5, no. 4, pp. 449–459, 2019.

[35] X. Yuan, B. Huang, Y. Wang, C. Yang, and W. Gui, "Deep learning-based feature representation and its application for soft sensor modeling with variable-wise weighted SAE," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 7, pp. 3235–3243, 2018.

[36] S. Guo, Y. Lin, S. Li, Z. Chen, and H. Wan, "Deep Spatial–Temporal 3D Convolutional Neural Networks for Traffic Data Forecasting," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, pp. 3913–3926, 2019.

[37] Y. Wu and H. Tan, "Short-term traffic flow forecasting with spatialtemporal correlation in a hybrid deep learning framework," 2016, https://arxiv.org/abs/1612.01022.

[38] D. Jo, B. Yu, H. Jeon, and K. Sohn, "Image-to-image learning to predict traffic speeds by considering area-wide spatiotemporal dependencies," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1188–1197, 2019.

[39] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *5th International Conference on Learning Representations (ICLR) 2017*, Toulon, France, April 2017.

[40] B. He, Z. Xu, Y. Xu, J. Hu, and Z. Ma, "Integrating Semantic Zoning Information with the Prediction of Road Link Speed Based on Taxi GPS Data," *Complexity*, vol. 2020, Article ID 6939328, 14 pages, 2020.

[41] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion Convolutional Recurrent Neural Network:Data-Driven Traffic Forecasting," in *Proceedings of the ICLR, Vancouver Convention Center*, Vancouver, BC, Canada, April 2018.

[42] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, "Graph Wavenet for Deep Spatial-Temporal Graph Modeling," in *Proceedings of the IJCAI*, pp. 1907–1913, Macao, China, August 2019.

[43] R. Huang, C. Huang, Y. Liu, G. Dai, and W. Kong, "Lsgcn: Long shortterm traffic prediction with graph convolutional networks," in *Proceedings of the International Joint Conferences on Artificial Intelligence Organization*, pp. 2355–2361, Yokohama, Japan, 2020.

[44] A. Roy, K. K. Roy, A. A. Ali, M. A. Amin, and A. M. Rahman, "Unified Spatio-Temporal Modeling for Traffic Forecasting Using Graph Neural Network," in *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, Shenzhen, China, July 2021.

[45] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR 2015*, San Diego, United States, May 2015.

[46] K. Xu, J. Ba, R. Kiros et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, pp. 2048–2057, Lille France, 2015.

[47] S. Li, X. Jin, Y. Xuan et al., "Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 5243–5253, Vancouver, BC, Canada, December 2019.

[48] D. Daiya and C. Lin, "Stock Movement Prediction and Portfolio Management via Multimodal Learning with Transformer," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021*, pp. 3305–3309, Toronto, ON, Canada, June 2021.

[49] H. Zhou, S. Zhang, J. Peng et al., "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35no. 12, pp. 11106–11115, Virtual Conference, February 2021.

[50] C. Park, C. Lee, H. Bahng, Y. Tae, S. Jin, and K. Kim, Eds.S. Ko and J. Choo, "ST-GRAT: A Novel Spatio-temporal Graph Attention Networks for Accurately Forecasting Dynamically Changing Road Speed," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20),*, p. 1215, New York, NY, USA, October 2020.

[51] X. Wang, Y. Ma, Y. Wang, W. Jin, X. Wang, J. Tang, C. Jia, and Y. Jian, Eds., "Traffic Flow Prediction Via Spatial Temporal Graph Neural Network," in *Proceedings of the Web Conference 2020 (WWW '20)*, pp. 1082–1092, New York, NY, USA, April 2020.

[52] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention Based Spatialtemporal Graph Convolutional Networks for Traffic Flow Forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 922–929, Honolulu, Hawaii, USA, 2019.

[53] J. You, R. Ying, and J. Leskovec, "Position-aware graph neural networks," *ICML'19*, vol. 97, pp. 7134–7143, 2019.

[54] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 914–921, New York Hilton Midtown, New York, New York, USA, 2020.

[55] M. Li and Z. Zhu, "Spatial-Temporal fusion graph neural networks for traffic flow forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35no. 5, pp. 4189–4196, Virtual Conference, 2021.

[56] Y. Chen, I. Segovia, and Y. R. Gel, "Z-GCNETs: Time Zigzags at Graph Convolutional Networks for Time Series Forecasting," in *Proceedings of the 38th International Conference on Machine Learning*, pp. 1684–1694, Virtual Conference, 2021.

[57] J. Choi, H. Choi, J. Hwang, and N. Park, "Graph neural controlled differential equations for traffic forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36no. 6, pp. 6367–6374, Virtual Conference, 2022.

[58] J. Yang, J. Li, W. Lu, L. Gao, and F. Mao, "Spatio-temporal DeepWalk Gated Recurrent Neural Network: A Deep Learning Framework for Traffic Learning and Forecasting," *Journal of Advanced Transportation*, vol. 2022, Article ID 4260244, 11 pages, 2022.