International Journal of
Geo-Information

Article

A3T-GCN: Attention Temporal Graph Convolutional Network for Traffic Forecasting

Jiandong Bai [1], Jiawei Zhu [2,*], Yujiao Song [3], Ling Zhao [2], Zhixiang Hou [4], Ronghua Du [4] and Haifeng Li [2]

[1] Beijing Institute of Tracking and Telecommunication Technology, Beijing 100094, China; baijiandongbv@gmail.com
[2] School of Geosciences and Info-Physics, Central South University, Changsha 410083, China; zhaolingg@csu.edu.cn (L.Z.); lihaifeng@csu.edu.cn (H.L.)
[3] Huawei Technologies Co., Ltd., Shenzhen 518129, China; songyujiao@huawei.com
[4] College of Automotive and Mechanical Engineering, Changsha University of Science and Technology, Changsha 410114, China; houzhixiang2008@163.com (Z.H.); csdrh@163.com (R.D.)
* Correspondence: jw_zhu@csu.edu.cn

Abstract: Accurate real-time traffic forecasting is a core technological problem against the implementation of the intelligent transportation system. However, it remains challenging considering the complex spatial and temporal dependencies among traffic flows. In the spatial dimension, due to the connectivity of the road network, the traffic flows between linked roads are closely related. In the temporal dimension, although there exists a tendency among adjacent time points, the importance of distant time points is not necessarily less than that of recent ones, since traffic flows are also affected by external factors. In this study, an attention temporal graph convolutional network (A3T-GCN) was proposed to simultaneously capture global temporal dynamics and spatial correlations in traffic flows. The A3T-GCN model learns the short-term trend by using the gated recurrent units and learns the spatial dependence based on the topology of the road network through the graph convolutional network. Moreover, the attention mechanism was introduced to adjust the importance of different time points and assemble global temporal information to improve prediction accuracy. Experimental results in real-world datasets demonstrate the effectiveness and robustness of the proposed A3T-GCN. We observe the improvements in RMSE of 2.51–46.15% and 2.45–49.32% over baselines for the SZ-taxi and Los-loop, respectively. Meanwhile, the Accuracies are 0.95–89.91% and 0.26–10.37% higher than the baselines for the SZ-taxi and Los-loop, respectively.

Keywords: traffic forecasting; attention temporal graph convolutional network; spatial dependence; temporal dependence

Citation: Bai, J.; Zhu, J.; Song, Y.; Zhao, L.; Hou, Z.; Du, R.; Li, H. A3T-GCN: Attention Temporal Graph Convolutional Network for Traffic Forecasting. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 485. https://doi.org/10.3390/ijgi10070485

Academic Editor: Wolfgang Kainz

Received: 28 May 2021
Accepted: 8 July 2021
Published: 15 July 2021

1. Introduction

Traffic forecasting is an important component of intelligent transportation systems and a vital part of transportation planning and management, and traffic control [1–4]. Accurate real-time traffic forecasting has been a great challenge because of complex spatiotemporal dependencies. Temporal dependence means that the traffic state changes with time, which is manifested by periodicity and tendency. Spatial dependence means that changes in traffic state are subject to the structural topology of road networks, which is manifested by the transmission of the upstream traffic state to downstream sections and the retrospective effects of the downstream traffic state on the upstream section [5]. Hence, considering the complex temporal features and the topological characteristics of the road network is essential for realizing the traffic forecasting task.

Existing traffic forecasting models can be divided into parametric models and non-parametric models. Common parametric models include historical average, time series [6,7], linear regression [8], and Kalman filtering models [9]. Although traditional parametric models use simple algorithms, they depend on stationary hypotheses. These models can neither reflect non-linearity and the uncertainty of traffic states nor overcome the interference of random events, such as traffic accidents. Non-parametric models can solve these problems well because they can learn the statistical laws of data automatically with adequate historical data. Common non-parametric models include k-nearest [10], support vector regression (SVR) [11,12], fuzzy logic [13], Bayesian network [14], and neural network models [15].

Recently, deep neural network models have attracted widespread attention from scholars because of the rapid development of deep learning [16–18]. Recurrent neural networks (RNNs), long short-term memory (LSTM) [19], and gated recurrent units (GRUs) [20] have been successfully utilized in traffic forecasting because they can use self-circulation mechanisms and model temporal dependence [21–23]. However, these models only consider the temporal variation of the traffic state and neglect spatial dependence. Many scholars have introduced convolutional neural networks (CNNs) in their models to characterize spatial dependence. Wu et al. [24] designed a feature fusion framework for short-term traffic flow forecasting by combining a CNN with LSTM. The framework captured the spatial characteristics of traffic flow through a one-dimensional CNN and explored short-term variations and the periodicity of traffic flow with two LSTMs. Cao et al. [25] proposed an end-to-end model called Interactive Temporal Recurrent Convolution Network (ITRCN), which transformed the interactive network flow to images and captured network flows using a CNN. ITRCN also extracted temporal features by using GRU. An experiment proved that the forecasting error of this method was 14.3% and 13.0% higher than those of GRU and CNN, respectively. Yu et al. [26] captured spatial correlation and temporal dynamics by the spatiotemporal recurrent convolutional networks (SRCN) based on deep convolutional neural network (DCNN) and LSTM. They also proved the superiority of SRCN based on the investigation of the traffic network data in Beijing. Sun et al. [27] proposed a deep-learning-based multi-branch model called Traffic Flow Forecasting Network (TFFNet) to forecast the short-term traffic flow. The TFFNet employed a multilayer fully convolutional framework to extract the hierarchical spatial dependencies from local to global scales.

Although CNN is applicable to Euclidean data [28], such as image and grids, it still has limitations in traffic networks, which possess non-Euclidean structures. In recent years, graph convolutional network (GCN) [29], which can overcome the abovementioned limitations and capture the structural characteristics of networks, has rapidly developed [30–32]. In addition, RNNs and their variants use sequential processing over time and are more apt to remember the latest information, thus are suitable for capturing evolving short-term tendencies. However, one problem with these models is that their performance will decrease as the prediction horizon increases. Moreover, the importance of different time points cannot be distinguished only by the proximity of time. As Pavlyuk suggested in [33], the definition of the spatiotemporal structure of traffic flow should also include relationships that are distant in time. Long-term dependencies (e.g., periodical dependencies) hidden in large time spans should be considered. Therefore, mechanisms that are capable of learning global correlations in long input time sequences are needed.

For this reason, an attention temporal GCN (A3T-GCN) was proposed for the traffic forecasting task. The A3T-GCN combines GCNs and GRUs and introduces an attention mechanism [34,35]. It can not only capture spatiotemporal dependencies but also adjust and assemble global variation information. The A3T-GCN is used for traffic forecasting on the basis of urban road networks.

The rest of the paper is organized as follows. Section 2 introduces the proposed model. In Section 3, we evaluate the performance of the A3T-GCN with real-world traffic datasets, including prediction results analysis, perturbation analysis, and visualization interpretation. We conclude the paper in Section 4.

2. Methods

2.1. Definition of Problems

In this study, traffic forecasting is performed to predict future traffic states according to historical traffic states on urban roads. Generally, traffic state can refer to traffic flow, speed, and density. In this study, traffic state only refers to traffic speed.

Definition 1. *Road network G: The topological structure of the urban road network is described as $G = (V, E)$, where $V = \{v_1, v_2, \cdots, v_N\}$ is the set of the road sections and N is the number of road sections. E is the set of edges, which reflects the connections between road sections. All of the connectivity information is stored in the adjacent matrix $A \in R^{N\times N}$, where rows and columns are indexed by road sections, and the value of each entry indicates the connectivity between corresponding road sections. The entry value is zero if there is no link between roads and one (unweighted graph) or non-negative (weighted graph) if otherwise.*

Definition 2. *Feature matrix $X^{N\times P}$: Traffic speed on a road section is viewed as the attribute of network nodes and it is expressed by the feature matrix $X \in R^{N\times P}$, where P is the number of node attribute features, that is, the length of the historical time series. X_i denotes the traffic speed in all sections at time i.*

Therefore, the traffic forecasting, modelling temporal and spatial dependencies, can be viewed as learning a mapping function f on the basis of the road network G and feature matrix X of the road network. Traffic speeds of future T moments are calculated as follows:

$$[X_{t+1}, \cdots, X_{t+T}] = f(G; (X_{t-n}, \cdots, X_{t-1}, X_t)), \tag{1}$$

where n is the length of a given historical time series and T is the length of time series that needs to be forecasted.

2.2. GCN Model

GCNs are semi-supervised models that can process graph structures. They are an advancement of CNNs in graph fields. GCNs have achieved much progress in many applications, such as image classification [36], document classification [28], and unsupervised learning [29]. Convolutional mode in GCNs includes spectrum and spatial domain convolutions [36]. The former was applied in this study. Spectrum convolution can be defined as the product of signal x on the graph and figure filter $g_\theta(L)$,which is constructed in the Fourier domain: $g_\theta(L) * x = Ug_\theta(U^T x)$, where θ is a model parameter, L is the graph Laplacian matrix, U is the eigenvector of normalized Laplacian matrix $L = I_N - D^{-\frac{1}{2}}AD^{-\frac{1}{2}} = U\lambda U^T$, and $U^T x$ is the graph Fourier transformation of x. x can also be promoted to $X \in R^{N\times C}$, where C refers to the number of features.

Given the characteristic matrix X and adjacent matrix A, GCNs can replace the convolutional operation in anterior CNNs by performing the spectrum convolutional operation with consideration of the graph node and the first-order adjacent domains of nodes to capture the spatial characteristics of the graph. Moreover, a hierarchical propagation rule is applied to superpose multiple networks. A multilayer GCN model in [29] is expressed as:

$$H^{(l+1)} = \sigma\left(\widetilde{D}^{-\frac{1}{2}}\,\widehat{A}\,\widetilde{D}^{-\frac{1}{2}}\,H^{(l)}\,\theta^{(l)}\right), \tag{2}$$

where $\widetilde{A} = A + I_N$ is an adjacent matrix with self-connection structures, I_N is an identity matrix, $\widetilde{D}$ is a degree matrix, $\widetilde{D}_{ii} = \sum_j \widetilde{A}_{ij}$, $H^{(l)} \in R^{N\times l}$ is the output of layer l, $\theta^{(l)}$ is the parameter of layer l, and $\sigma(\cdot)$ is an activation function used for nonlinear modeling.

Generally, a two-layer GCN model [29] can be expressed as:

$$f(X, A) = \sigma\left(\widehat{A}\,ReLU\left(\widehat{A}\,X\,W_0\right)W_1\right), \tag{3}$$

where X is a feature matrix; A is the adjacent matrix; and $\widehat{A} = \widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}$ is a preprocessing step, where $\widetilde{A} = A + I_N$ is the adjacent matrix of graph G with a self-connection structure. $W_0 \in R^{P\times H}$ is the weight matrix from the input layer to the hidden unit layer, where P is the length of time and H is the number of hidden units. $W_1 \in R^{H\times T}$ is the weight matrix from the hidden layer to the output layer. $f(X, A) \in R^{N\times T}$ denotes the output with a forecasting length of T, and $ReLU()$ is a common nonlinear activation function.

GCNs can encode the topological structures of road networks and the attributes of road sections simultaneously by determining the topological relationship between the central road section and the surrounding road sections. Spatial dependence can be captured on this basis. In a word, this study learned spatial dependence through the GCN model [29].

2.3. GRU Model

Temporal dependence of traffic state is another key problem that hinders traffic forecasting. RNNs are neural network models that process sequential data. However, limitations in long-term forecasting are observed in traditional RNNs because of disadvantages in terms of gradient disappearance and explosion [37]. LSTM [19] and GRUs [20] are variants of RNNs that mediate the problems effectively. LSTM and GRUs have basically the same fundamental principles. Both models use gated mechanisms to maintain long-term information and perform similarly on various tasks [38]. Compared with GRUs, LSTM has an additional memory cell and adapts more gating units to control the information flow. Thus, GRU has a relatively simpler structure, fewer parameters, and is easier to compute and implement [39].

In the present model, temporal dependence was captured by a GRU model. The calculation process is introduced as follows, where h_{t-1} is the hidden state at $t-1$, X_t is the traffic speed at the current moment, and r_t is the reset gate to control the degree of neglecting the state information at the previous moment. Information irrelevant to forecasting can be abandoned. If the reset gate outputs 0, then the traffic information at the previous moment is neglected. If the reset gate outputs 1, then the traffic information at the previous moment is brought into the next moment completely. u_t is the update gate and is used to control the state information quantity at the previous moment that is brought into the current state. Meanwhile, c_t is the memory content stored at the current moment, and h_t is the output state at the current moment.

$$u_t = \sigma(W_u * [X_t, h_{t-1}] + b_u) \tag{4}$$

$$r_t = \sigma(W_r * [X_t, h_{t-1}] + b_r) \tag{5}$$

$$c_t = \tanh(W_c[X_t, (r_t * h_{t-1})] + b_c) \tag{6}$$

$$h_t = u_t * h_{t-1} + (1 - u_t) * c_t. \tag{7}$$

GRUs determine the traffic state at the current moment by using the hidden state at the previous moment and traffic information at the current moment as input. GRUs retain the variation trends of historical traffic information when capturing traffic information at the current moment because of the gated mechanism. Hence, this model can capture dynamic temporal variation features from the traffic data, that is, this study has applied a GRU model to learn the temporal variation trends of the traffic state.

2.4. Attention Model

The attention model is realized on the basis of the encoder–decoder model. This model was initially used in neural machine translation tasks [40]. Nowadays, attention models are widely applied to image caption generation [34], recommendation systems [41] and document classification [42]. With the rapid development of such models, existing attention models can be divided into multiple types, such as soft and hard attention [40], global and local attention [43], and self-attention [35]. In the current study, a soft attention model was used to learn the importance of traffic information at every moment, and then a context vector that could express the global variation trends of the traffic state was calculated for future traffic forecasting tasks.

Suppose that a time series $X_i(i = 1, 2, \cdots, n)$, where n is the time series length, is introduced. The design process of soft attention models is introduced as follows. First, the hidden states $h_i(i = 1, 2, \cdots, n)$ at different moments are calculated using CNNs (and their variants) or RNNs (and their variant), and they are expressed as $H = \{h_1, h_2, \cdots, h_n\}$. Second, a scoring function is designed to calculate the score/weight of each hidden state. Third, an attention function is designed to calculate the context vector C_t that can describe global traffic variation information. Finally, the final output results are obtained using the context vector. In the present study, these steps were followed in the design process, but a multilayer perception was applied as the scoring function instead.

Particularly, the characteristics h_i at each moment were used as input when calculating the weight of each hidden state based on the attention mechanism. The corresponding outputs could be gained through two hidden layers. The weights of each characteristic α_i were calculated by a Softmax normalized index function (Equation (8)), where $w_{(1)}$ and $b_{(1)}$ are the weight and deviation of the first layer and $w_{(2)}$ and $b_{(2)}$ are the weight and deviation of the second layer, respectively.

$$e_i = w_{(2)}(w_{(1)}H + b_{(1)}) + b_{(2)} \tag{8}$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_{k=1}^{n} \exp(e_k)}. \tag{9}$$

Finally, the attention function was designed. The calculation process of the context vector C_t that covers global traffic variation information is shown in Equation (10).

$$C_t = \sum_{i=1}^{n} \alpha_i * h_i. \tag{10}$$

2.5. A3T-GCN Model

The A3t-GCN is an improvement of our previous work named the Temporal Graph Convolutional Network (T-GCN) [31]. The attention mechanism was introduced to re-weight the influence of historical traffic states and thus to capture the global variation trends of the traffic state. The model structure is shown in Figure 1.

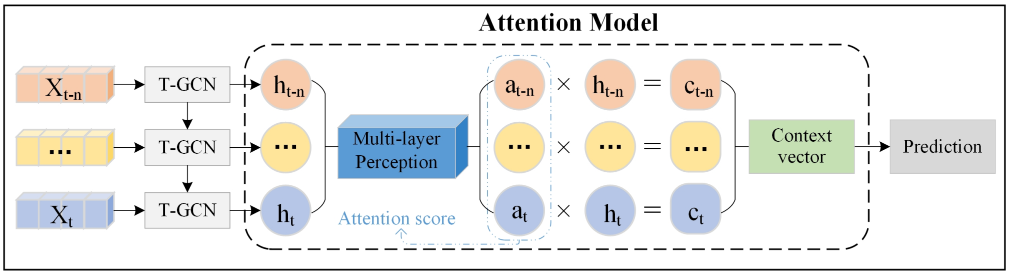


Figure 1. A3T-GCN framework.

The T-GCN model was constructed by combining GCN and GRU. First, n time steps of the historical traffic data were input into the T-GCN model to obtain n hidden states (h) that covered spatiotemporal characteristics: $\{h_{t-n}, \cdots, h_{t-1}, h_t\}$. The calculation of the T-GCN is shown in Equations (11)–(14) [31], where h_{t-1} is the output at $t - 1$. GC is the graph convolutional process. u_t and r_t are the update and reset gates at t, respectively. c_t is the stored content at the current moment. h_t is the output state at moment t, and W and b are the weight and the deviation in the training process, respectively.

$$u_t = \sigma(W_u * [GC(A, X_t), h_{t-1}] + b_u) \tag{11}$$

$$r_t = \sigma(W_r * [GC(A, X_t), h_{t-1}] + b_r) \tag{12}$$

$$c_t = \tanh(W_c * [GC(A, X_t), (r_t * h_{t-1})] + b_c) \tag{13}$$

$$h_t = u_t * h_{t-1} + (1 - u_t) * c_t). \tag{14}$$

Then, the hidden states were fed into the attention model to determine the context vector that covers the global traffic variation information. Particularly, the weight of each h was calculated by Softmax using a multilayer perception: $\{a_{t-n}, \cdots, a_{t-1}, a_t\}$. The context vector that covers global traffic variation information was calculated by the weighted sum. Finally, forecasting results were outputted using the fully connected layer.

In sum, we proposed the A3T-GCN to realize traffic forecasting. The urban road network was constructed into a graph network and the traffic state on different sections was described as node attributes. The topological characteristics of the road network were captured by a GCN to obtain spatial dependence. The dynamic variation of node attributes was captured by a GRU to obtain the local temporal tendency of the traffic state. The global variation trend of the traffic state was then captured by the attention model, which was conducive to realizing accurate traffic forecasting.

2.6. Loss Function

Training aims to minimize errors between real and predicted speed in the road network. The real and predicted speeds on different sections at t are expressed by Y and $\widehat{Y}$, respectively. Therefore, the objective function of A3T-GCN is shown as follows. The first term aims to minimize the error between real and predicted speed. The second term L_{reg} is a normalization term, which is conducive to avoid overfitting. λ is a hyper-parameter.

$$loss = \parallel Y_t - \widehat{Y_t} \parallel + \lambda L_{reg}. \tag{15}$$

3. Experiments

3.1. Data Description

Two real-world traffic datasets, namely, the taxi trajectory dataset (SZ-taxi) in Shenzhen City and the loop detector dataset (Los-loop) in Los Angeles, were used. Both datasets are related to the traffic speed. Hence, traffic speed is viewed as the traffic information in the experiments. The SZ-taxi dataset is the taxi trajectory of Shenzhen from 1 January to 31 January 2015. In the present study, 156 major roads of Luohu District were selected as the study area, so the size of the SZ-taxi adjacency matrix is 156 × 156. The Los-loop dataset was collected on the highway of Los Angeles County in real-time by loop detectors. A total of 207 sensors, along with their traffic speeds from 1 March to 7 March 2012, were selected. The size of the Los-loop adjacency matrix is thus 207 × 207.

3.2. Evaluation Metrics

To evaluate the prediction performance of the model, the error between real traffic speed and predicted results is evaluated based on the metrics following [31]:

(1) Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{MN}\sum_{j=1}^{M}\sum_{i=1}^{N}(y_i^j - \widehat{y_i^j})^2}. \tag{16}$$

(2) Mean Absolute Error (MAE):

$$MAE = \frac{1}{MN}\sum_{j=1}^{M}\sum_{i=1}^{N}\left|y_i^j - \widehat{y_i^j}\right|. \tag{17}$$

(3) Accuracy:

$$Accuracy = 1 - \frac{\| Y - \widehat{Y} \|_F}{\| Y \|_F}, \tag{18}$$

where F is the Frobenius norm, which calculates the square root of the sum of the absolute squares of elements in the matrix.

(4) Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum_{j=1}^{M} \sum_{i=1}^{N} (y_i^j - \widehat{y_i^j})^2}{\sum_{j=1}^{M} \sum_{i=1}^{N} (y_i^j - \bar{Y})^2}. \tag{19}$$

(5) Explained Variance Score (var):

$$var = 1 - \frac{Var\left\{Y - \widehat{Y}\right\}}{Var\{Y\}}, \tag{20}$$

where y_i^j and $\widehat{y_i^j}$ are the real and predicted traffic information of temporal sample j on road i, respectively. N is the number of nodes on the road. M is the number of temporal samples. Y and $\widehat{Y}$ are the set of y_i^j and $\widehat{y_i^j}$ respectively, and $\bar{Y}$ is the mean of Y.

Particularly, RMSE and MAE were used to measure prediction error. Small RMSE and MASE values reflect high prediction precision. Accuracy is used to measure forecasting precision, and a high accuracy value is preferred. R^2 and var calculate the correlation coefficient, which measures the ability of the prediction result to represent the actual data: the larger the value is, the better the prediction effect is.

3.3. *Experimental Result Analysis*

The hyper-parameters of A3T-GCN include learning rate, epoch and the number of hidden units. In the experiment, learning rate and epoch were manually set, on the basis of experiences, as 0.001 and 5000 for both datasets, and parameters were randomly initialized from a normal distribution. We tested the model on different numbers of hidden units (8, 16, 32, 64, 100, 128) and chose the numbers that reported the best performance, that is, 64 for the SZ-taxi dataset and 100 for the Los-loop dataset.

In the present study, 80% of the traffic data were used as the training set, and the remaining 20% of the data were used as the test set. The traffic information in the next 15, 30, 45, and 60 min was predicted. The predicted results were compared with results from the historical average model (HA), the auto-regressive integrated moving average model (ARIMA), SVR, the GCN model, and the GRU model. The A3T-GCN was analyzed from the perspectives of precision, spatiotemporal prediction capabilities, long-term prediction capability and global feature capturing capability.

3.3.1. High Prediction Precision

Tables 1 and 2 show the comparisons of different models and two real datasets in terms of the prediction precision of various traffic speed lengths. We bold the optimal results in tables, and $*$ means that the values are small enough to be negligible, indicating that the model's prediction effect is poor. The prediction precision of neural network models (e.g., A3T-GCN and GRU) is higher than that of other models (e.g., HA, ARIMA, and SVR). With respect to the 15-min time series, the RMSE and accuracy of HA are approximately 9.22% higher and 4.24% lower than those of A3T-GCN, respectively. The RMSE and accuracy of ARIMA are approximately 46.15% higher and 39.01% lower than those of A3T-GCN, respectively. The RMSE and accuracy of SVR are approximately 5.95% higher and 2.81% lower than those of A3T-GCN, respectively. Compared with GRU, the RMSE and accuracy of HA are approximately 6.88% higher and 3.32% lower, respectively. The RMSE and accuracy of ARIMA are approximately 44.76% and 38.07%, respectively.

The RMSE and accuracy of SVAR are approximately 3.52% and 1.87%, respectively. These results are mainly caused by the poor nonlinear fitting abilities of HA, ARIMA, and SVAR with regard to complicated, changing traffic data. Processing long-term non-stationary data is difficult when ARIMA is used. Moreover, ARIMA gains by averaging the errors of different sections. The data of some sections might greatly fluctuate to increase the final error. Hence, ARIMA shows the lowest forecasting accuracy.

Table 1. The prediction results of the A3T-GCN model and other baseline methods on the SZ-taxi dataset.

T	Metric	HA	ARIMA	SVR	GCN	GRU	A3T-GCN
15 min	RMSE	4.2951	7.2406	4.1455	5.6596	3.9994	**3.8989**
	MAE	2.7815	4.9824	2.6233	4.2367	**2.5955**	2.6840
	Accuracy	0.7008	0.4463	0.7112	0.6107	0.7249	**0.7318**
	R^2	0.8307	*	0.8423	0.6654	0.8329	**0.8512**
	var	0.8307	0.0035	0.8424	0.6655	0.8329	**0.8512**
30 min	RMSE	4.2951	6.7899	4.1628	5.6918	4.0942	**3.9228**
	MAE	2.7815	4.6765	**2.6875**	4.2647	2.6906	2.7038
	Accuracy	0.7008	0.3845	0.7100	0.6085	0.7184	**0.7302**
	R^2	0.8307	*	0.8410	0.6616	0.8249	**0.8493**
	var	0.8307	0.0081	0.8413	0.6617	0.8250	**0.8493**
45 min	RMSE	4.2951	6.7852	4.1885	5.7142	4.1534	**3.9461**
	MAE	2.7815	4.6734	2.7359	4.2844	2.7743	**2.7261**
	Accuracy	0.7008	0.3847	0.7082	0.6069	0.7143	**0.7286**
	R^2	0.8307	*	0.8391	0.6589	0.8198	**0.8474**
	var	0.8307	0.0087	0.8397	0.6590	0.8199	**0.8474**
60 min	RMSE	4.2951	6.7708	4.2156	5.7361	4.0747	**3.9707**
	MAE	2.7815	4.6655	2.7751	4.3034	2.7712	**2.7391**
	Accuracy	0.7008	0.3851	0.7063	0.6054	0.7197	**0.7269**
	R^2	0.8307	*	0.8370	0.6564	0.8266	**0.8454**
	var	0.8307	0.0111	0.8379	0.6564	0.8267	**0.8454**

* Means that the values are small enough to be negligible, indicating that the model's prediction effect is poor.

Table 2. The prediction results of the A3T-GCN model and other baseline methods on the Los-loop dataset.

T	Metric	HA	ARIMA	SVR	GCN	GRU	A3T-GCN
15 min	RMSE	7.4427	10.0439	6.0084	7.7922	5.2182	**5.0904**
	MAE	4.0145	7.6832	3.7285	5.3525	**3.0602**	3.1365
	Accuracy	0.8733	0.8275	0.8977	0.8673	0.9109	**0.9133**
	R^2	0.7121	0.0025	0.8123	0.6843	0.8576	**0.8653**
	var	0.7121	*	0.8146	0.6844	0.8577	**0.8653**
30 min	RMSE	7.4427	9.3450	6.9588	8.3353	6.2802	**5.9974**
	MAE	4.0145	7.6891	3.7248	5.6118	**3.6505**	3.6610
	Accuracy	0.8733	0.8275	0.8815	0.8581	0.8931	**0.8979**
	R^2	0.7121	0.0031	0.7492	0.6402	0.7957	**0.8137**
	var	0.7121	*	0.7523	0.6404	0.7958	**0.8137**
45 min	RMSE	7.4427	10.0508	7.7504	8.8036	7.0343	**6.6840**
	MAE	**4.0145**	7.6924	4.1288	5.9534	4.0915	4.1712
	Accuracy	0.8733	0.8273	0.8680	0.8500	0.8801	**0.8861**
	R^2	0.7121	*	0.6899	0.5999	0.7446	**0.7694**
	var	0.7121	0.0035	0.6947	0.6001	0.7451	**0.7705**
60 min	RMSE	7.4427	10.0538	8.4388	9.2657	7.6621	**7.0990**
	MAE	**4.0145**	7.6952	4.5036	6.2892	4.5186	4.2343
	Accuracy	0.8733	0.8273	0.8562	0.8421	0.8694	**0.8790**
	R^2	0.7121	*	0.6336	0.5583	0.6980	**0.7407**
	var	0.7121	0.0036	0.5593	0.5593	0.6984	**0.7415**

* Means that the values are small enough to be negligible, indicating that the model's prediction effect is poor.

Similar conclusions could be drawn for Los-loop. The A3T-GCN model can obtain the optimal prediction performance of all metrics in two real datasets, thus proving the validity and superiority of the A3T-GCN model in spatiotemporal traffic forecasting tasks.

3.3.2. Effectiveness of Modeling Both Spatial and Temporal Dependencies

To test the benefits of depicting the spatiotemporal characteristics of traffic data simultaneously in A3T-GCN, the A3T-GCN model is compared with GCN and GRU.

Figure 2 shows the results based on SZ-taxi. The prediction error of A3T-GCN is kept lower than that of GCN (considering spatial characteristics only) in 15, 30, 45, and 60 min of traffic forecasting, as shown in Figure 2a. Compared with GCN, A3T-GCN achieves approximately 31.11%, 31.08%, 30.94%, and 30.78% lower RMSEs in 15, 30, 45, and 60 min of the traffic forecasting time series, respectively. Compared with GRU (considering temporal characteristics only), A3T-GCN achieves approximately 2.51% lower RMSE in 15 min traffic forecasting, approximately 4.19% lower RMSE in 30 min traffic forecasting, approximately 4.99% lower RMSE in the 45 min time series, and approximately 2.55% lower RMSE in the 60 min time series. In sum, the prediction error of A3T-GCN is kept lower than that of GRU in 15, 30, 45, and 60 min traffic forecasting, as shown in Figure 2b.

Results based on Los-loop are similar to those based on SZ-taxi. The prediction error of A3T-GCN is kept lower than that of GCN in 15, 30, 45, and 60 min of traffic forecasting, as shown in Figure 3a. A3T-GCN achieves approximately 34.67%, 28.05%, 24.08%, and 23.38% lower RMSEs in 15, 30, 45, and 60 min of traffic forecasting time series, respectively. As for the comparison with GRU shown in Figure 3b, A3T-GCN achieves approximately 2.45% lower RMSE in 15 min traffic forecasting, approximately 4.50% lower RMSE in 30 min traffic forecasting, approximately 4.98% lower RMSE in the 45 min time series, and approximately 7.35% lower RMSE in the 60 min time series.

Figures 2 and 3 show that the A3T-GCN has better prediction capabilities than the GCN and the GRU, respectively. In other words, the A3T-GCN model can capture the spatial topological characteristics of urban road networks and the temporal variation characteristics of the traffic state simultaneously, which means it maintains superiority under various prediction horizons compared with the GRU and the GCN.

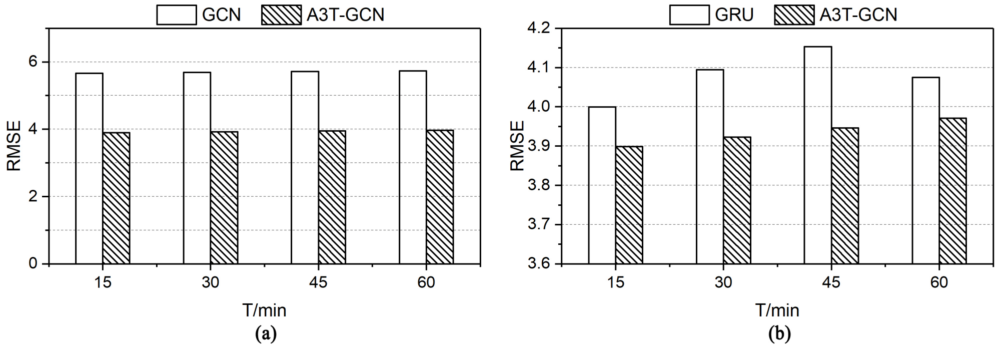


Figure 2. SZ-taxi: Effectiveness of modeling both spatial and temporal dependencies. (**a**) The RMSE comparison between the A3T-GCN and the GCN (captures spatial features only). The outperformance of the A3T-GCN under various prediction horizons verifies the effectiveness of the A3T-GCN in capturing additional temporal features compared with the GCN. (**b**) The RMSE comparison between the A3T-GCN and the GRU (captures temporal features only). The outperformance of the A3T-GCN under various prediction horizons verifies the effectiveness of the A3T-GCN in capturing additional spatial features compared with GRU.

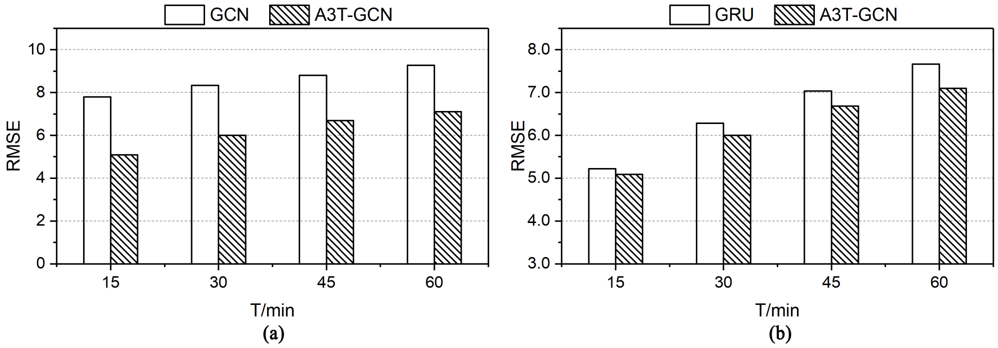


Figure 3. Los-loop: Effectiveness of modeling both spatial and temporal dependencies. (**a**) The RMSE comparison between the A3T-GCN and the GCN (captures spatial features only). The outperformance of the A3T-GCN under various prediction horizons verifies the effectiveness of the A3T-GCN in capturing additional temporal features compared with the GCN. (**b**) The RMSE comparison between the A3T-GCN and the GRU (captures temporal features only). The outperformance of the A3T-GCN under various prediction horizons verifies the effectiveness of the A3T-GCN in capturing additional spatial features compared with GRU.

3.3.3. Long-Term Prediction Capability

The long-term prediction capability of A3T-GCN was tested for traffic speed forecasting in 15, 30, 45, and 60 min prediction horizons. The RMSE comparison of different models under different prediction horizons is shown in Figure 4a. The RMSE of the A3T-GCN is the lowest under all lengths of time series. The variation trends of RMSE and accuracy, which reflect prediction error and precision, respectively, of the A3T-GCN under different prediction horizons, are shown in Figure 4b. We can observe that the RMSE increases as the length of the time series increases, whereas the accuracy declines; both change slightly and show a certain stability. Compared with GRU (which outperforms other baselines), A3T-GCN has a standard deviation of approximately 0.03 in RMSE, while that of the GRU is approximately 0.06.

The forecasting results based on Los-loop are shown in Figure 5, and consistent laws are found. In sum, A3T-GCN has good long-term prediction capability. It can maintain the best performance under 15, 30, 45, and 60 min prediction horizons, as shown in Figure 5a. Figure 5b shows that the forecasting results of A3T-GCN change slightly with changes in the prediction horizons, thereby showing a certain stability. Compared with GRU (which outperforms other baselines), A3T-GCN has a standard deviation of approximately 0.76 in RMSE, while that of the GRU is approximately 0.91. Therefore, the A3T-GCN is applicable to short-term and long-term traffic forecasting tasks.

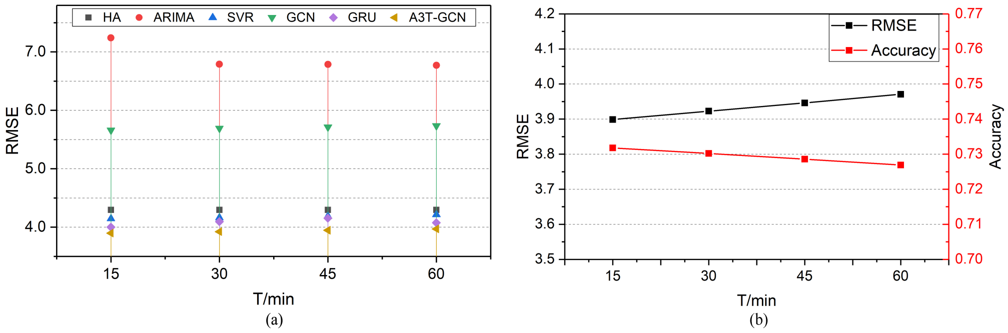


Figure 4. SZ-taxi: Long-term prediction ability. (**a**) The RMSE of the A3T-GCN and baselines under different prediction horizons. (**b**) The changes in RMSE and accuracy of the A3T-GCN under different prediction horizons.

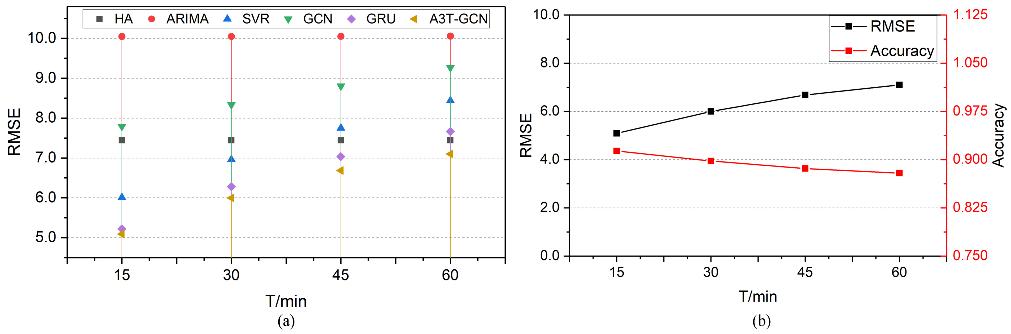


Figure 5. Los-loop: Long-term prediction ability. (**a**) The RMSE of the A3T-GCN and baselines under different prediction horizons. (**b**) The changes in RMSE and accuracy of the A3T-GCN under different prediction horizons.

3.3.4. Effectiveness of Introducing Attention to Capture Global Variation

A3T-GCN and T-GCN were compared to test the superiority of capturing global variation. The results are shown in Table 3. The A3T-GCN model shows approximately 0.86% lower RMSE and approximately 0.32% higher accuracy than the T-GCN model under the 15 min time series; approximately 1.31% lower RMSE and approximately 0.48% higher accuracy under the 30 min time series; approximately 1.14% lower RMSE and approximately 0.43% higher accuracy under 45 min traffic forecasting; and approximately 0.99% lower RMSE and approximately 0.37% higher accuracy under the 60 min time series.

Hence, the prediction error of A3T-GCN is lower than that of T-GCN, but the accuracy of the former is higher under different horizons of traffic forecasting, thereby proving the global feature capturing capability of the A3T-GCN model.

Table 3. Comparison of forecasting results between A3T-GCN and T-GCN under different lengths of time series based on SZ-taxi and Los-loop.

T	Metric	SZ-Taxi		Los-Loop	
		T-GCN	A3T-GCN	T-GCN	A3T-GCN
15 min	*RMSE*	3.9325	3.8989	5.1264	5.0904
	MAE	2.7145	2.6840	3.1802	3.1365
	Accuracy	0.7295	0.7318	0.9127	0.9133
	R^2	0.8539	0.8512	0.8634	0.8653
	var	0.8539	0.8512	0.8634	0.8653
30 min	*RMSE*	3.9740	3.9228	6.0598	5.9974
	MAE	2.7522	2.7038	3.7466	3.6610
	Accuracy	0.7267	0.7302	0.8968	0.8979
	R^2	0.8451	0.8493	0.8098	0.8137
	var	0.8451	0.8493	0.8100	0.8137
45 min	*RMSE*	3.9910	3.9461	6.7065	6.684
	MAE	2.7645	2.7261	4.1158	4.1712
	Accuracy	0.7255	0.7286	0.8857	0.8861
	R^2	0.8436	0.8474	0.7679	0.7694
	var	0.8436	0.8474	0.7684	0.7705
60 min	*RMSE*	4.0099	3.9707	7.2677	7.099
	MAE	2.7860	2.7391	4.6021	4.2343
	Accuracy	0.7242	0.7269	0.8762	0.8790
	R^2	0.8421	0.8454	0.7283	0.7407
	var	0.8421	0.8454	0.7290	0.7415

3.4. *Perturbation Analysis*

Noise is inevitable in real-world datasets. Therefore, perturbation analysis was conducted to test the robustness of A3T-GCN. In this experiment, two types of random noises were added to the traffic data. Random noise obeys the Gaussian distribution $N \in (0, \sigma^2)$, where $\sigma \in (0.2, 0.4, 0.8, 1, 2)$, and the Poisson distribution $P(\lambda)$, where $\lambda \in (1, 2, 4, 8, 16)$. The noise matrix values were normalized to [0,1].

The experimental results based on SZ-taxi are shown in Figure 6. The results of adding Gaussian noise are shown in Figure 6a, where the x- and y-axes show the changes in σ and in different evaluation metrics, respectively. Different colors represent various metrics. Similarly, the results of adding Poisson noise are shown in Figure 6b. The values of different evaluation metrics remain basically the same regardless of the changes in σ/λ. Hence, the proposed model can, remarkably, resist noise.

Figure 7a,b indicates that the experimental results based on Los-loop are consistent with the experimental results based on SZ-taxi. Therefore, the A3T-GCN model can remarkably resist noise and still obtain stable forecasting results under Gaussian and Poisson perturbations.

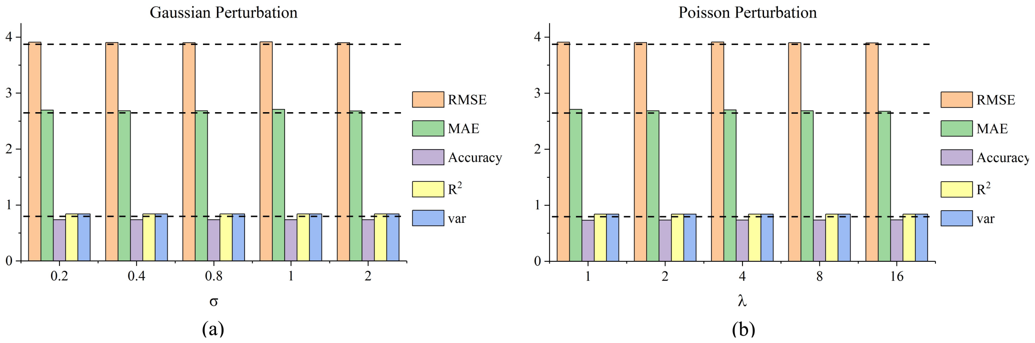


Figure 6. SZ-taxi: Perturbation analysis. (**a**) The performances of the A3T-GCN after adding Gaussian perturbation to the SZ-taxi dataset. (**b**) The performances of the A3T-GCN after adding Poisson perturbation to the SZ-taxi dataset.

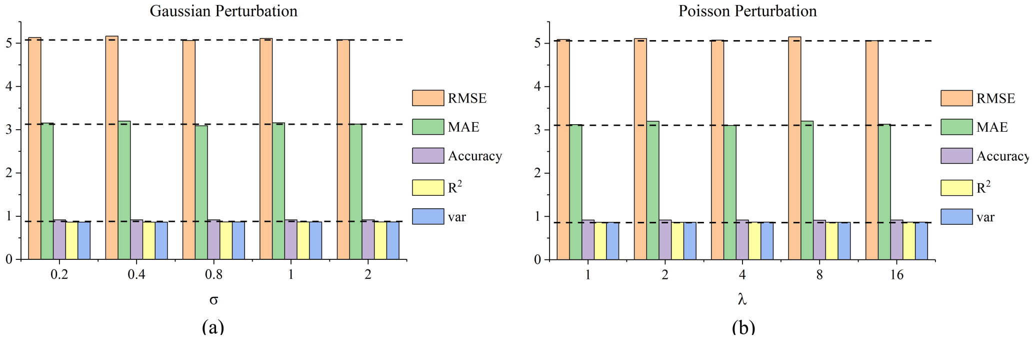


Figure 7. Los-loop: Perturbation analysis. (**a**) The performances of the A3T-GCN after adding Gaussian perturbation to the Los-loop dataset. (**b**) The performances of the A3T-GCN after adding Poisson perturbation to the Los-loop dataset.

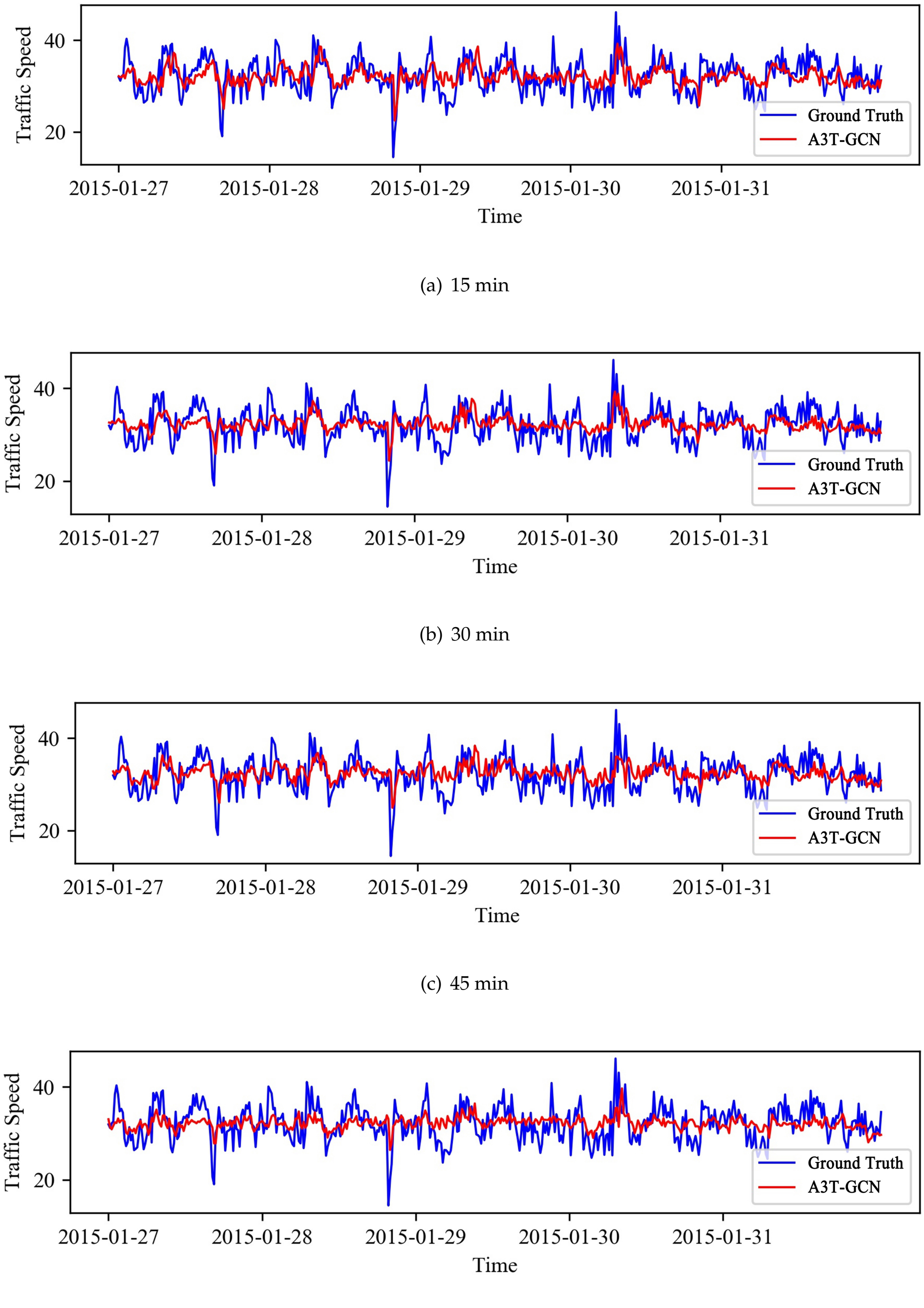
3.5. Visualized Analysis

The forecasting results of the A3T-GCN model based on two real datasets are visualized for a good explanation of the model.

(1) SZ-taxi: We visualize the result of one road on 27 January 2015. Visualization results for the 15, 30, 45, and 60 min time series are shown in Figure 8.

(2) Los-loop: Similarly, we visualize one loop detector data in the Los-loop dataset. Visualization results for the 15, 30, 45, and 60 min are shown in Figure 9.

In sum, the predicted traffic speed shows a similar variation trend to the actual traffic speed under different time series lengths, which suggests that the A3T-GCN model is competent at the traffic forecasting task. This model can also capture the variation trends of traffic speed and recognize the start and end points of rush hours. As suggested by the well-fit of A3T-GCN to the abrupt drops during rush hours in Figure 9, the A3T-GCN model forecasts traffic jams accurately, thereby proving its validity in real-time traffic forecasting.



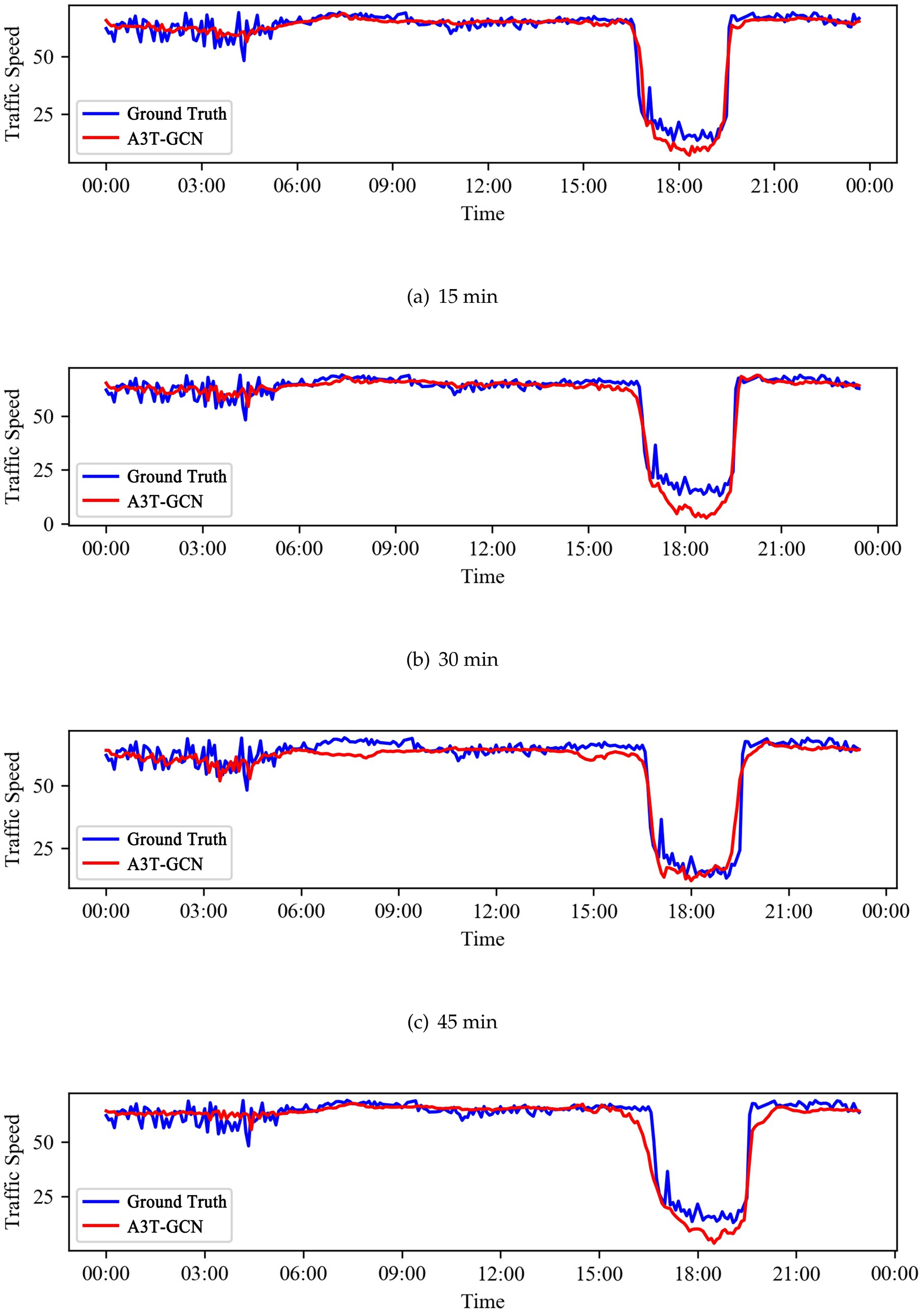
(a) 15 min

(b) 30 min

(c) 45 min

(d) 60 min

Figure 8. The visualization results for prediction horizon of 15, 30, 45, 60 min (SZ-taxi).



(a) 15 min

(b) 30 min

(c) 45 min

(d) 60 min

Figure 9. The visualization results for prediction horizon of 15, 30, 45, 60 min (Los-loop).

4. Conclusions

A traffic forecasting method called A3T-GCN is proposed to capture global temporal dynamics and spatial correlations simultaneously and to facilitate traffic forecasting. The urban road network is constructed into a graph, and the traffic speed on the roads is described as attributes of nodes on the graph. In the proposed method, the spatial dependencies are captured by GCN based on the topological characteristics of the road network. Meanwhile, the dynamic variation of the sequential historical traffic speeds is captured by GRU. Moreover, the global temporal variation trend is captured and assembled by the attention mechanism. Finally, the proposed A3T-GCN model is tested on the urban road network-based traffic forecasting task using two real datasets, namely, SZ-taxi and Los-loop. We observe the improvements in RMSE of 2.51–46.15% and 2.45–49.32% over baselines for the SZ-taxi and Los-loop, respectively. Meanwhile, the accuracies are 0.95–89.91% and 0.26–10.37% higher than the baselines for the SZ-taxi and Los-loop, respectively. The results show that the A3T-GCN model is superior to HA, ARIMA, SVR, GCN, GRU, and T-GCN in terms of prediction precision under different lengths of prediction horizon, thereby proving its validity for real-time traffic forecasting.

Author Contributions: Conceptualization, Jiandong Bai and Jiawei Zhu; methodology, Yujiao Song; software, Jiandong Bai and Yujiao Song; data curation, Ling Zhao, Zhixiang Hou and Ronghua Du; writing—original draft preparation, Jiandong Bai and Jiawei Zhu; visualization, Yujiao Song; supervision, Jiawei Zhu and Haifeng Li. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant numbers 41571397, 41871364); and the Fundamental Research Funds (No.2019zzts881) for the Central Universities of Central South University.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Huang, H.J. Dynamic Modeling of Urban Transportation Networks and Analysis of Its Travel Behaviors. *Chin. J. Manag.* **2005**, *2*, 18–22.
2. Jian, Y.; Fan, B. Synthesis of Short-Term Traffic Flow Forecasting Research Progress. *Urban Transp. China* **2012**, *10*, 73–79.
3. Jing, L.; Wei, G. A Summary of Traffic Flow Forecasting Methods. *J. Highw. Transp. Res. Dev.* **2004**, *3*, 82–85.
4. Gao, L.; Liu, X.; Liu, Y.; Wang, P.; Deng, M.; Zhu, Q.; Li, H. Measuring road network topology vulnerability by ricci curvature. *Phys. A Stat. Mech. Its Appl.* **2019**, *527*, 121071. [CrossRef]
5. Dong, C.J.; Shao, C.F.; Zhuge, C.X.; Meng, M. Spatial and Temporal Characteristics for Congested Traffic on Urban Expressway. *J. Beijing Univ. Technol.* **2012**, *38*, 1242–1246, 1268.
6. Ahmed, M.S.; Cook, A.R. *Analysis of Freeway Traffic Time-Series Data by Using Box-Jenkins Techniques*; Transportation Research Record: Washington, DC, USA, 1979.
7. Hodge, V.J.; Krishnan, R.; Austin, J.; Polak, J.; Jackson, T. Short-term prediction of traffic flow using a binary neural network. *Neural Comput. Appl.* **2014**, *25*, 1639–1655. [CrossRef]
8. Sun, H.; Zhang, C.; Ran, B. Interval prediction for traffic time series using local linear predictor. In Proceedings of the International IEEE Conference on Intelligent Transportation Systems, Washington, WA, USA, 3–6 October 2004.
9. Okutani, I.; Stephanedes, Y.J. Dynamic prediction of traffic volume through Kalman filtering theory. *Transp. Res. Part B Methodol.* **1984**, *18*, 1–11. [CrossRef]
10. Altman, N.S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992**, *46*, 175–185.
11. Wu, C.H.; Ho, J.M.; Lee, D. Travel-Time Prediction With Support Vector Regression. *IEEE Trans. Intell. Transp. Syst.* **2005**, *5*, 276–281. [CrossRef]
12. Fu, G.; Han, G.; Lu, F.; Xu, Z. Short-Term Traffic Flow Forecasting Model Based on Support Vector Machine Regression. *J. South China Univ. Technol.* **2013**, *41*, 71–76.
13. Yin, H.; Wong, S.C.; Xu, J.; Wong, C.K. Urban traffic flow prediction using a fuzzy-neural approach. *Transp. Res. Part C* **2002**, *10*, 85–98. [CrossRef]
14. Sun, S.; Zhang, C.; Yu, G. A bayesian network approach to traffic flow forecasting. *IEEE Trans. Intell. Transp. Syst.* **2006**, *7*, 124–132. [CrossRef]
15. Çetiner B.G.; Sari, M.; Borat, O. A Neural Network Based Traffic-Flow Prediction Model. *Math. Comput. Appl.* **2010**, *15*, 269–278.
16. Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A. Mastering the game of Go without human knowledge. *Nature* **2017**, *550*, 354–359. [CrossRef] [PubMed]
17. Morav?ík, M.; Schmid, M.; Burch, N.; Lisy, V.; Morrill, D.; Bard, N.; Davis, T.; Waugh, K.; Johanson, M.; Bowling, M. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science* **2017**, *356*, eaam6960.
18. Yuan, H.; Li, G. A Survey of Traffic Prediction: from Spatio-Temporal Data to Intelligent Transportation. *Data Sci. Eng.* **2021**, *6*, 63–85. [CrossRef]
19. Graves, A. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780.
20. Cho, K.; Merrienboer, B.V.; Bahdanau, D.; Bengio, Y. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *arXiv* **2014**, arXiv:1409.1259.

21. Rui, F.; Zuo, Z.; Li, L. Using LSTM and GRU neural network methods for traffic flow prediction. In Proceedings of the Youth Academic Conference of Chinese Association of Automation, Wuhan, China, 11–13 November 2016
22. Lint, J.W.C.V.; Hooqendoorn, S.P.; Zuvlen, H.J.V. Freeway Travel Time Prediction with State-Space Neural Networks: Modeling State-Space Dynamics with Recurrent Neural Networks. *Transp. Res. Rec. J. Transp. Res. Board* **2002**, *1811*, 347–369.
23. Zhao, F.; Zeng, G.Q.; Lu, K.D. EnLSTM-WPEO: Short-term traffic flow prediction by ensemble LSTM, NNCT weight integration, and population extremal optimization. *IEEE Trans. Veh. Technol.* **2019**, *69*, 101–113. [CrossRef]
24. Wu, Y.; Tan, H. Short-term traffic flow forecasting with spatial-temporal correlation in a hybrid deep learning framework. *arXiv* **2016**, arXiv:1612.01022.
25. Cao, X.; Zhong, Y.; Yun, Z.; Jiang, W.; Zhang, W. Interactive Temporal Recurrent Convolution Network for Traffic Prediction in Data Centers. *IEEE Access* **2017**, *6*, 5276–5289. [CrossRef]
26. Yu, H.; Wu, Z.; Wang, S.; Wang, Y.; Ma, X. Spatiotemporal Recurrent Convolutional Networks for Traffic Prediction in Transportation Networks. *Sensors* **2017**, *17*, 1501. [CrossRef] [PubMed]
27. Sun, S.; Wu, H.; Xiang, L. City-Wide Traffic Flow Forecasting Using a Deep Convolutional Neural Network. *Sensors* **2020**, *20*, 421. [CrossRef] [PubMed]
28. Defferrard, M.; Bresson, X.; Vandergheynst, P. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 3844–3852.
29. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv* **2016**, arXiv:1609.02907.
30. Li, Y.; Yu, R.; Shahabi, C.; Liu, Y. Graph Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. *arXiv* **2017**, arXiv:1707.01926.
31. Zhao, L.; Song, Y.; Zhang, C.; Liu, Y.; Wang, P.; Lin, T.; Deng, M.; Li, H. T-gcn: A temporal graph convolutional network for traffic prediction. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 3848–3858. [CrossRef]
32. Yu, B.; Lee, Y.; Sohn, K. Forecasting road traffic speeds by considering area-wide spatio-temporal dependencies based on a graph convolutional neural network (GCN). *Transp. Res. Part C Emerg. Technol.* **2020**, *114*, 189–204. [CrossRef]
33. Pavlyuk, D. Feature selection and extraction in spatiotemporal traffic forecasting: A systematic literature review. *Eur. Transp. Res. Rev.* **2019**, *11*, 1–19. [CrossRef]
34. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *arXiv* **2015**, arXiv:cs.LG/1502.03044.
35. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Long Beach, CA, USA, 2017; pp. 5998–6008.
36. Bruna, J.; Zaremba, W.; Szlam, A.; Lecun, Y. Spectral Networks and Locally Connected Networks on Graphs. *arXiv* **2013**, arXiv:1312.6203.
37. Bengio, Y.; Simard, P.; Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **2002**, *5*, 157–166. [CrossRef]
38. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**, arXiv:cs.NE/1412.3555.
39. Cho, K.; Merrienboer, B.V.; Gulcehre, C.; BaHdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv* **2014**, arXiv:1406.1078.
40. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv* **2014**, arXiv:cs.CL/1409.0473.
41. Xiao, J.; Ye, H.; He, X.; Zhang, H.; Wu, F.; Chua, T.S. Attentional Factorization Machines: Learning the Weight of Feature Interactions via Attention Networks. *arXiv* **2017**, arXiv:cs.LG/1708.04617.
42. Pappas, N.; Popescu-Belis, A. Multilingual Hierarchical Attention Networks for Document Classification. *arXiv* **2017**, arXiv:1707.00896.
43. Luong, M.T.; Pham, H.; Manning, C.D. Effective Approaches to Attention-based Neural Machine Translation. *arXiv* **2015**, arXiv:1508.04025.