DOI: 10.1111/mice.12688

ORIGINAL ARTICLE

COMPUTER-AIDED CIVIL AND INFRASTRUCTURE ENGINEERING WILEY

Spatiotemporal gated graph attention network for urban traffic flow prediction based on license plate recognition data

Jinjun Tang | **Jie Zeng**

Smart Transport Key Laboratory of Hunan Province, School of Traffic and Transportation Engineering, Central South University, Changsha, China

Correspondence
Jinjun Tang, Smart Transport Key Laboratory of Hunan Province, School of Traffic and Transportation Engineering, Central South University, Changsha, 410075, China.
Email: jinjuntang@csu.edu.cn

Funding information
National Key R&D Program of China (No. 2020YFB1600400), the Natural Science Foundation of Hunan Province (No. 2020JJ4752), Innovation-Driven Project of Central South University (No. 2020CX041), Foundation of Central South University (No. 502045002), National Training Program of Innovation and Entrepreneurship for Undergraduates (2020105330040).

Abstract

The accurate forecasting of traffic states is an essential application of intelligent transportation system. Due to the periodic signal control at intersections, the traffic flow in an urban road network is often disturbed and expresses intermittent features. This study proposes a forecasting framework named the spatiotemporal gated graph attention network (STGGAT) model to achieve accurate predictions for network-scale traffic flows on urban roads. Based on license plate recognition (LPR) records, the average travel times and volume transition relationships are estimated to construct weighted directed graphs. The proposed STGGAT model integrates a gated recurrent unit layer, a graph attention network layer with edge features, a gated mechanism based on the bidirectional long short-term memory and a residual structure to extract the spatiotemporal dependencies of the approach- and lane-level traffic volumes. Validated on the LPR system in Changsha, China, STGGAT demonstrates superior accuracy and stability to those of the baselines and reveals its inductive learning and fault tolerance capabilities.

1 | INTRODUCTION

The intelligent transportation system (ITS) has been considered a countermeasure for handling severe traffic problems in cities, such as traffic jams and air pollution. As a critical technology in the ITS, the reliable and accurate prediction of future traffic states is an essential component for the release of traffic information, route guidance, traffic management optimization (Hashemi & Abdelghany, 2018), and greenhouse gas emissions volume prediction (Ganji et al., 2020). Traffic flow prediction on an urban road network is a significant technology in the ITS, and it is valuable for both traffic managers and travelers. For traffic managers, traffic state perception and prediction are playing an increasingly vital role in traffic management and control. For travelers, accurate traffic prediction can provide real-time traffic information, which is useful in travel guidance for avoiding congestion. Furthermore, compared with traffic flow prediction on a highway network, the accurate prediction of traffic flows on urban road networks is much more difficult due to complicated environments and interference, such as signal timing and the interference caused by vehicles entering and leaving the main road. Hence, traffic flow prediction is a vital and significant research topic to be addressed in traffic engineering. Facing a massive amount of traffic flow data collected from various detectors, it becomes a significant challenge to explore the dynamic characteristics of traffic flows and conduct accurate forecasting for urban road networks.

© 2021 *Computer-Aided Civil and Infrastructure Engineering*

Since researchers first applied traffic flow prediction models on the freeway in the 1970s (Prigogine et al., 1972), this field has attracted increasing research attention, and numerous emerging technologies have been developed to improve prediction performance. Generally, existing studies focusing on traffic flow prediction fall into two categories (Y. Zhang, Cheng, et al., 2019): model- and data-driven studies. Based on comprehensive prior knowledge, including queuing theory (L. Li & Chen, 2013) and traffic flow theory (Adeli & Ghosh-Dastidar, 2004; Y. Zhang, Smirnova, et al., 2018), model-driven approaches establish simulation models to describe traffic flow distributions and the decision-making processes of drivers. These established models can reconstruct traffic conditions in the real world and explore traffic states under different environmental situations by adjusting their parameters. However, the states of traffic flow can be affected by a large number of factors, such as extreme weather, special events, drivers' unique characteristics, and experience, leading to unreasonable analysis results. In addition, due to the different topological structures of road networks, barriers exist to transferring these simulated models to other situations.

With the rapid development of data collection technologies, data-driven approaches have become increasingly popular due to their capabilities to explore the characteristics and patterns of historical traffic flow data. These methods focus on mining the regularities hidden in historical data to predict future traffic states instead of modeling the dynamic behavior evolution process in the simulated traffic system (Zhao et al., 2019). Currently, data-driven approaches can be classified into two categories: Parametric methods and nonparametric methods.

For parametric models, the structure of a model is predetermined by several theoretical assumptions, and the parameters are estimated from historical data (Tang, Li, et al., 2019). Due to their stable performances and convenient calculations, parametric approaches attracted researchers' interests in an early stage of this field. Among all the parametric methods, the autoregressive integrated moving average (ARIMA) model and its extended models, such as the seasonal ARIMA (Williams & Hoel, 2003), subset ARIMA (Lee & Fambro, 1999), and hybrid empirical mode decomposition (EMD)-ARIMA (H. Wang et al., 2016), are improved approaches concerning the application of time-series analysis for traffic prediction. In addition, other approaches, including partial least squares (W. Li et al., 2020), Kalman filtering (Kumar, 2017; Okutani & Stephanedes, 1984), and generalized autoregressive conditional heteroscedasticity (GARCH; Y. Zhang et al., 2015), are also widely used in traffic flow prediction. Min et al. (2009) established a hybrid model combining the spatiotemporal ARIMA with a dynamic turn ratio prediction model to improve the prediction accuracy of the model at urban intersections. Zou et al. (2015) combined the space-time model, vector autoregression, and ARIMA to forecast traffic speeds on a freeway. J. Guo et al. (2014) proposed an adaptive Kalman filter to update the process variances and utilized it to implement a combination of the stochastic ARIMA and GARCH. In these methods, the subsequent predictions are determined by prior forecasts. Hence, the prediction error accumulates during the multistep prediction task. Furthermore, as these models are dependent on the assumption that the future distribution of traffic flow expresses a similar pattern to those of historical distributions, the prediction performance suffers sharp declines under intense fluctuations in traffic flow data. To fill this gap, a large number of nonparametric approaches have been proposed to improve the prediction performances of such models. The conventional methods used in traffic flow prediction include artificial neural networks (Jiang & Adeli, 2004; Jiang et al., 2005; L. Li et al., 2019; Tang et al., 2017), support vector machines (Feng et al., 2019; Tang, Chen, et al., 2019; Yao et al., 2017), random forests (Hamner, 2010), and *k*-nearest neighbors (kNNs; L. Zhang et al., 2013). Without the underlying stationary assumptions, nonparametric models can fit different traffic conditions and capture the inherent relationship between the previous information and future traffic status.

More recently, with the advances in computing performance, many advanced deep learning models have been widely applied for traffic flow prediction because of their advantages of strong learning ability, accurate and stable prediction performance, and deep feature extraction. Huang et al. (2014) proposed a deep architecture by stacking a deep belief network and a multitask regression layer, and this was the first time deep learning approaches were applied in this field. Lv et al. (2014) established an autoencoder model to represent generic traffic flow features and achieved superior performance to those of traditional models. Due to the unique module of the self-circulation mechanism, recurrent neural network (RNN)-based models express advantages in temporal evolution mining. Ma et al. (2015) employed a long short-term memory (LSTM) network to capture the nonlinear dynamics in time-series traffic data. Gu et al. (2019) constructed a two-layer RNN model consisting of LSTM and gated recurrent unit (GRU) for lane-level speed prediction. Since the convolutional neural network (CNN) has a powerful ability to capture the spatial relationships between grids, it is suitable to deal with traffic flow prediction for a massive road network. To apply CNNs in traffic flow prediction, most studies (L. Liu et al., 2019; Y. Liu et al., 2020; J. Zhang, Zheng, et al., 2018; J. Zhang, Zheng, et al., 2020; Zheng et al., 2020) need to divide the study area into grids first, and then treat each grid as a pixel of an image. One

exception is the model of Dai et al. (2019) who represented the sensor network as an image and proposed an algorithm to rearrange the order of sensors in the image according to the correlation coefficients between them.

In addition to the temporal evolution of traffic states, there exist apparent spatial correlations between traffic flows at different sections of the road network (Ermagun & Levinson, 2019). Hence, effectively capturing spatial dependencies can contribute to traffic flow prediction accuracy. However, in traditional RNN models, the spatial relationships of traffic flows are difficult to extract, while CNN-based models are limited to Euclidean-structured data and are inapplicable for extracting the topological characteristics hidden in traffic networks. To integrate road network topology into the prediction model, the graph neural network (GNN) was introduced to capture the spatiotemporal correlations between network-scale traffic flows in recent years. Y. Li et al. (2017) proposed a hybrid model to capture the spatiotemporal correlations on a directed graph, where a random walk was utilized to represent the spatial relationship, and an encoder-decoder architecture was employed to capture the temporal dependencies. Zhang, Cheng, et al. (2019) constructed a spatial–temporal graph inception residual network) for traffic speed prediction on a large-scale directed graph and adopted a residual learning process and an inception module to enhance the prediction performance of the network. AGC-Seq2Seq, proposed by Z. Zhang, Li, et al. (2019), utilizes the Seq2Seq structure and a graph convolutional network (GCN) to model spatial and temporal correlations separately and incorporates an attention mechanism to overcome the shortcomings of multistep speed prediction. However, the aforementioned GCN-based models encounter limitations in directed graphs and inductive learning tasks. A capable GCN variant, the graph attention network (GAT; Veličković et al., 2017), relies on its ability to update the importance of neighbor nodes automatically and is also applied in traffic flow prediction (Pan et al., 2019; Park et al., 2020; C. Zhang, Yu, et al., 2019). Compared with the widely-used GCN framework, the GAT enhances the capability of the network to take advantage of the structures of directed graphs. Moreover, the GAT shows a powerful capability to conduct inductive learning, while GCNs cannot be applied to graphs with different topology structures.

In summary, although numerous advanced traffic flow prediction models have been developed in recent decades, there still exist several challenges that have not been addressed, and we summarize them as follows:

1. Due to signal control at intersections, traffic flows on urban roads are intermittently interrupted, leading to fluctuations at short-term scales. The high-accuracy prediction of traffic flows at the network scale is a great challenge in traffic management. Although a few researchers have studied this field (Do et al., 2019; Wu et al., 2016), there are still gaps in the literature regarding effective methods for predicting approach- and lane-level traffic volumes from a network-scale perspective.
2. Existing methods tend to select the adjacent relationship of the physical road network (Cheng et al., ; Cui et al., 2020; Guo et al., 2019; Zhang,et al., 2019; Zhao et al., 2019) or employ the distance among all the nodes to construct graphs for graph neural networks (Pan et al., 2019; Park et al., 2020; Zheng et al., 2020). However, both the adjacency and distances are static factors. It is difficult to reflect the actual traffic distributions and characteristics in the road network. For instance, to avoid traffic jams, drivers may be more likely to select routes with longer distances but lower travel times. Thus, it is a challenge to reflect the characteristics of network-scale traffic flows in extracted topology graphs.
3. Existing GCN models always assume fixed spatial correlations between roads, ignoring the dynamic dependencies, and thus they cannot achieve inductive learning on different graph structures. Furthermore, for GAT-based models, few studies have considered a weighted directed graph and the temporal dependencies of previous traffic states.

In this study, we propose a prediction framework for approach- and lane-level traffic volumes at network-scale urban intersections called the spatiotemporal gated GAT (STGGAT). First, we extract the average travel time (ATT) and volume transition matrices among all the approaches or lanes based on license plate recognition (LPR) records and then adopt the complex network construction method proposed by Cupertino et al. (2013) to establish a weighted directed graph, where the nodes represent detectors and edges denote connection relationships. In the prediction model, GRU is utilized to replace the linear transformation in the naive GAT and capture the temporal dependencies of previous traffic data; we also propose an RNN-based gated module to determine the importance of each head in the multi-head attention mechanism. Moreover, edge features are introduced into the self-attention aggregator of the GAT as prior knowledge to improve the spatial dependencies. The primary contributions of this paper can be summarized as follows:

1. We study network-scale traffic flow prediction at the approach and lane levels based on widely equipped LPR devices. Different from the distribution of traffic flows on freeways, the density of an urban road

network is relatively high, and there exist complicated spatial correlations between traffic flows at different sections. Furthermore, traffic flows collected at urban intersections are more intermittent and intense than those on freeways due to periodic interruptions from signal control. These fluctuation characteristics at a short-term scale pose a great challenge with regard to forecasting.

2. We develop an urban road network transformation method to transform a physical road network into a weighted directed graph. Extracted from the LPR records, the ATT is treated as the distance measure, and the volume transition relationships are utilized to reflect the interactions among intersection approaches or lanes in the road network. An improved complex network construction method is applied to establish a weighted directed graph considering the ATT and volume transition. Since the graph construction measure is data-driven, it can reflect the traffic distribution and the characteristics of the studied road network more comprehensively.
3. A STGGAT is proposed to achieve network-scale traffic flow prediction at the approach and lane levels. Based on the constructed weighted directed graph, the edge features are integrated into the GAT model, and we employ GRU to extract the temporal evolution of previous traffic states. Moreover, we utilize the bidirectional LSTM (BiLSTM) to explore the dependencies between different heads in the multi-head attention mechanism.
4. A validation of the traffic flow prediction performance of this model is conducted based on the data collected from LPR devices on the urban road network in Changsha, China, and we compare the proposed STGGAT with other widely used models. The STGGAT model exhibits high prediction accuracy and stability at both the approach and lane levels. In addition, the STGGAT model also exhibits superior performance in the conducted fault tolerance analysis and strong capability for inductive learning tasks.

The remainder of this paper is organized as follows. The weighted directed graph construction method and the problem formulation for traffic flow prediction are introduced in Section 2. Section 3 describes the STGGAT model and its components in detail. Section 4 presents the data description utilized in this study. Section 5 compares the performance of the proposed model with those of other widely used baselines and explores the stability and robustness of the models. Finally, we draw a conclusion about this study and provide an outlook for future researches in Section 6.

2 | PRELIMINARIES

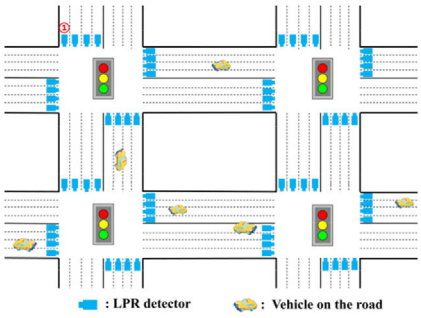
2.1 | Weighted directed graph construction

An urban road network can be viewed as a weighted directed graph $G = (V, E, A, W)$, where $V, E, A \in R^{N\times N}$, and $W \in R^{N\times N}$ denote a set of N nodes, a set of M edges, the adjacency matrix, and the weight matrix, respectively. In the traffic road network, the nodes can represent the detectors, and the edges denote the connections between different detectors. Specifically, since LPR devices can collect traffic volumes at different levels, the nodes can represent the traffic volumes at the lane, approach, or intersection levels.

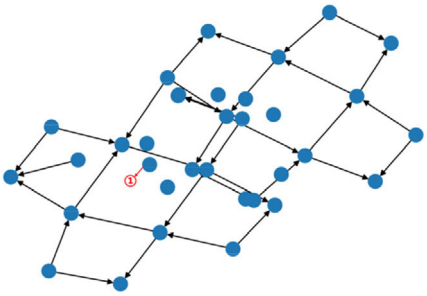
Naturally, it is easy to consider establishing a traffic graph directly based on the connection relationships of the road network. In several previous studies, traffic states were estimated from taxi trajectory data, and graphs were constructed based on adjacent relationships (Y. Zhang, Cheng, et al., 2019; Z. Zhang, Li et al., 2019; Zhao et al., 2019). However, limited by the location accuracy of GPS equipment, it is challenging to obtain lane-level traffic states. Furthermore, there are several obstacles when constructing a directed graph based on connection relationships and LPR records:

1. In a road network with high LPR equipment coverage, as illustrated in an example in Figure 1(a), all the lanes in these four intersections can be transformed into the directed graph displayed in Figure 1(b).[1] However, not all intersections in the city are equipped with LPR detectors. Hence, defining the connection relationships in an area with low detector coverage, such as in Figure 1(c), is still a critical challenge.
2. Due to traffic control measures, vehicles in different lanes have different directions when going through intersections, including going straight, turning left, turning right, and going in a combination of directions. To construct a directed graph based on connections, the traffic direction and control measures of each lane must be investigated accurately first.
3. In an urban road network, not all adjacent roads have strong spatial correlations. Hence, if the graph representation is only constructed based on the physical road network, this may yield several useless connections, resulting in poor prediction accuracy.

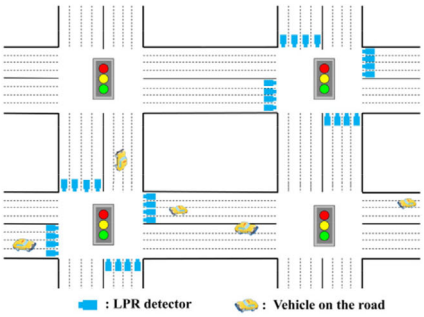
[1] The isolated nodes in Figure 1(b) represent specific lanes on the boundary, such as "1" in Figure 1(a). Due to traffic rules, vehicles traveling in this lane cannot reach any nodes with LPR detectors in the selected area, nor can any vehicle traveling in another lane reach them.



(a) road network with high LPR equipment coverage



(b) topology structure of the physical road network



(c) road network with relatively low LPR equipment coverage

FIGURE 1 Transform the road network to a directed graph based on the physical topology

Transforming a road network to a topology graph based on its physical structure requires the consideration that traffic states are positively correlated with nearby traffic flows due to the movements of vehicles between different sections of the road network. Vehicles traveling on roads are most likely to access adjacent roads within a short distance, representing "proximity in space." In addition to the adjacent relationships, researchers have also utilized the distances among all the nodes to construct the highway network (Pan et al., 2019; Park et al., 2020; Zheng et al., 2020). From the perspective of "proximity in time," the travel time between roads that are "proximal in space" is relatively lower than that between nonadjacent roads. However, both the adjacent relationships and the distances are static factors, and they cannot be affected by the traffic characteristics of the road network. In real-world traffic systems, the effects of different road infrastructure levels and traffic conditions may lead to a nonproportional relationship between the travel time and road length. To a certain extent, "proximity in time" can reflect "proximity in space," which means travel time can express the connection relationships. Essentially, these two graph construction methods, adjacent relationships and distances, can be regarded as connections between nodes that are "proximal in time." In detail, if two roads are adjacent, the travel time between them is relatively short; conversely, if there is a short travel time between two roads, then the two roads are likely to be adjacent. Cui et al. (2019) considered the impact of traffic transmission between nonadjacent road segments and proposed a free-flow reachable matrix to ensure the rationality of the physical graph. In this study, we employ the travel times and volume transition relationships between different sections in the road network to construct a weighted directed graph.

In the following subsections, we first extract the ATT and volume transition matrices from the LPR records, and then employ the complex network construction method proposed by Cupertino et al. (2013) to transform the urban road network into a weighted directed graph.

2.1.1 | Extraction of traffic features

From the LPR records, we can obtain the collected sections and the timestamp of each vehicle going through the intersections. Hence, the vehicle trajectory during one trip can be extracted as follows:

$$\begin{cases} L^m = \{l_1^m,\ l_2^m, \dots,\ l_n^m\} \\ T^m = \{t_1^m,\ t_2^m, \dots,\ t_n^m\} \\ t_{i+1}^m - t_i^m < t_\varepsilon,\ i = 1, 2, \dots, n-1 \end{cases} \tag{1}$$

Here, L^m denotes the set of intersections that vehicle m continuously passes through, t_i^m represents the collection time of vehicle m at intersection i, and t_s denotes the time threshold that is used to divide the itinerary. In this study, t_s is set to 20 min.

Generally, travel time represents the time interval between a specified origin and destination (Dharia & Adeli, 2003; Ghosh-dastidar et al., 2006). In this study, we estimate the travel time between every two records during one trip. In detail, for a trip with p intersections, there are $\binom{p}{2}$ pairs of travel times between different sections. Suppose $ATT \in R^{N\times N}$, $VTM \in R^{N\times N}$ denote the ATT and volume transition matrices, respectively; then, the proposed algorithm is summarized in Algorithm 1, and the traffic feature extraction process is shown in Figure 2.

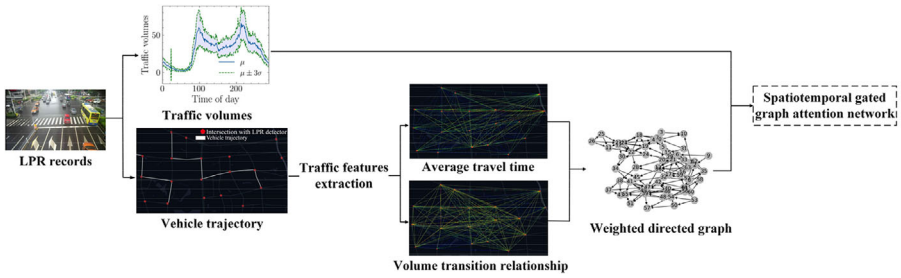


FIGURE 2 The framework of the road network transformation method

ALGORITHM 1. Traffic features extraction.

Input: Trajectory set, $L=\{L^1,L^2,...,L^m\}$, responding travel time set, $T=\{T^1,T^2,...,T^m\}$.

Output: Travel time matrix, ATT, volume transition matrix, VTM.

for all vehicles $i\in[1,m]$ **do**
  Extract the start point l_1^i and end point l_n^i.
  for the passing-by collection locations $j\in[1,n-1]$ **do**
    for the passing-by collection locations $k\in[2,n]$
      $VTM(l_j^i,l_k^i)\leftarrow VTM(l_j^i,l_k^i)+1$
      $ATT(l_j^i,l_k^i)\leftarrow ATT(l_j^i,l_k^i)+t_k^i-t_j^i$
    end for
  end for
end for
$ATT\leftarrow ATT./VTM$

2.1.2 | Road network transformation

In the existing studies related to traffic flow prediction based on detector data, methods for transforming road networks into traffic graphs can be divided into two basic approaches (Belkin & Niyogi, 2003): (1) Each node, donating a detector, is connected to its k nearest nodes, named the kNNs. (2) Each node is connected to all nodes within a specific distance, named the ε-radius neighbors. It is noted that the aforementioned "distance" can be both the distance in space and the correlation coefficient between different detectors. However, the constructed networks based on these methods may be too dense or sparse and not strongly connected, leading to isolated nodes.

In the field of complex networks, Cupertino et al. (2013) combined the aforementioned kNNs and ε-radius neighbors methods and proposed an adaptive approach to construct a strongly connected and relatively sparse network based on the single-linkage method (Sibson, 1973). In this study, we adopt this algorithm to transform the road network into a weighted directed graph based on the ATT matrix ATT and volume transition matrix VTM. The framework of the road network transformation method developed in this study is displayed in Algorithm 2.

ALGORITHM 2. Network construction method.

Input: Average travel time matrix, $ATT\in R^{N\times N}$; hyperparameters, k, λ.

Output: Directed graph G;

Create a non-connected network including N nodes.
Put each node i into a unique node group G_i.
while the number of group > 1 **do**
  Identify the nearest two node groups, denoted as G_m and G_n.
  Calculate the average dissimilarity inside G_m and G_n, denoted as D_m and D_n.
  Select the nearest k pairs of nodes between G_m and G_n into node pair set γ.
  $d_c\leftarrow\lambda\cdot\max(D_m,D_n)$.
  for each node pair $(\vartheta_1^i,\vartheta_2^i)\subseteq\gamma$ **do**
    if $ATT(\vartheta_i^1,\vartheta_i^2)<d_c$ **then**
      Add a directed edge from ϑ_1^i to ϑ_2^i.
    end if
  end for
  Merge connected nodes into a node group.
  Update the distance between each two node groups.
  Update the number of the remaining groups.
end while

Furthermore, because future traffic volumes are profoundly affected by previous traffic states (Y. Zhang, Cheng, et al., 2019), we add a self-loop to each node. After the directed traffic network is constructed, to extend the established graph to a weighted directed graph, we utilize the volume transition matrix to represent the edge weights. For the edge from nodes i to j, the weight E_{ij} can be calculated as follows:

$$E_{ij}=\frac{VTM(i,j)}{\sum_{k\in\mathbb{N}_i}VTM(i,k)} \tag{2}$$

where $\mathbb{N}_i$ denotes the set of all destinations beginning from node i. Supposing that the historical traffic data of each node may have the most influential effect on its future traffic states, we set the weights of the self-loops to 1. As mentioned above, for adjacent roads in the physical road network, not only is the traffic state on adjacent roads

proximal in space and time, but the transition volumes are also relatively numerous. Similarly, the edge weights can achieve importance-based sampling between the target road and the adjacent roads, where edges with larger transition volumes are regarded as more vital.

2.2 | Problem formulation

Generally, network-scale traffic flow prediction aims at exploring the spatiotemporal dependencies hidden in previous traffic data to predict future traffic volumes. Assuming x_t^i represents the traffic volumes collected from detector i at time t, the traffic flow prediction problem can be summarized as follows: Given the historical traffic data of the road network with N detectors $X_t = \{X_t^1, X_t^2, \dots, X_t^N\}$ in the previous s time steps (where $X_t^i = \{x_{t-s+1}^i, x_{t-s+2}^i, \dots, x_t^i\}$) and a specific graph structure G, we aim to learn a mapping function f for estimating the future traffic volumes $y_{t+1} = \{x_{t+1}^1, x_{t+1}^2, \dots, x_{t+1}^N\}$:

$$f : (G; X_t) \rightarrow y_{t+1} \tag{3}$$

3 | METHODOLOGY

3.1 | Framework of the prediction model

Figure 3 illustrates the architecture of the proposed STGGAT model for short-term traffic flow prediction on urban road networks. In summary, the STGGAT model consists of four components: A GRU layer, a GAT layer, an RNN-based gated mechanism, and a residual structure. First, GRU is utilized to improve the ability of the model to capture the temporal correlations in the historical traffic data. Then, we incorporate the edge features into the naive GAT model to enhance the spatial dependencies between different sections in the constructed weighted directed graphs. The RNN-based gated module is employed to determine the importance of different heads in the multi-head attention mechanism. Moreover, we adopt the residual connection structure to accelerate the training process and improve the convergence efficiency. Each component module of the proposed STGGAT model is described in the following subsections in detail.

3.2 | Spatial dependencies

In this subsection, we employ the GAT to capture the spatial dependencies in the urban road network. This type of

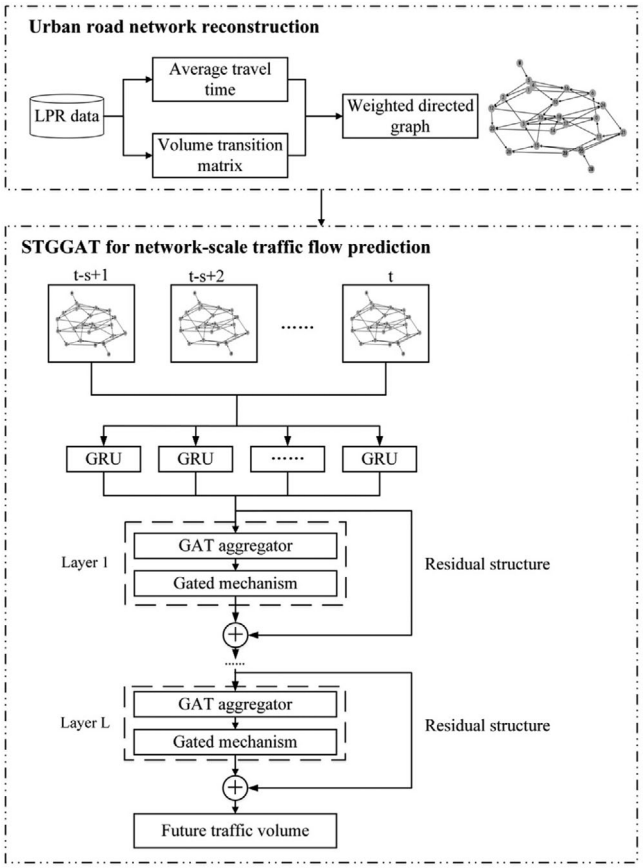


FIGURE 3 The spatiotemporal learning framework for network-scale traffic flow prediction

network has achieved success in numerous tasks, including computer version (Bao et al., 2019) and recommendation systems (X. Wang, He, et al., 2019). Compared with the GCN model, which is widely used in traffic flow prediction, there are several superiorities inherent in the GAT: (1) It can achieve the assignment of different weights to different neighbors, leading to advantages in dealing with directed graphs. (2) The model parameters are related to the features of each node instead of the structure of the graph. Hence, the GAT can be applied to inductive learning tasks. In other words, we can utilize the model trained on a specific road network to predict future traffic states on other road networks.

Assume the input features of node i at layer l in the GAT are denoted as $h_i^{(l)}$. Specifically, $h_i^{(1)}$ represents the input previous traffic states X_t^i, which are defined in Section 2.2. The detailed formulation of the GAT (Veličković et al., 2017) at layer l is introduced as follows:

$$z_i^{(l)} = W^{(l)} h_i^{(l)} \tag{4}$$

$$e_{ij}^{(l)} = \text{LeakyReLU}(\vec{a}^{\mathrm{T}}[z_i^{(l)} || z_j^{(l)}]) \tag{5}$$

$$\alpha_{ij}^{(l)} = \mathrm{softmax}(e_{ij}^{(l)}) = \frac{\exp(e_{ij}^{(l)})}{\sum_{k\in\mathbb{N}_i}\exp(e_{ik}^{(l)})} \tag{6}$$

$$h_i^{(l+1)} = \sigma\Big(\sum_{j\in\mathbb{N}_i}\alpha_{ij}^{(l)} z_j^{(l)}\Big) \tag{7}$$

$$\mathrm{LeakyReLU}(x) = \begin{cases} x, & x \geq 0 \\ \alpha \cdot x, & x < 0 \end{cases} \tag{8}$$

Here, $W^{(l)}$ is a learnable weight matrix that aims to increase the dimensionality of the input, $\vec{a}^T$ is a single-layer feedforward neural network that acts as a self-attention mechanism, $||$ represents the concatenation operation, and $\sigma(\cdot)$ represents the activation function for applying a nonlinearity. In this study, the negative input scope of LeakyReLU adopts the setting ($\alpha = 0.2$) as in Veličković et al. (2017).

Moreover, to enhance the optimization stability during the training process of the self-attention mechanism, a multi-head attention mechanism is applied by the following equation:

$$h_i^{(l+1)} = \begin{cases} \sigma\Big(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in\mathbb{N}_i}\alpha_{ij,k}^{(l)} W_k^{(l)} h_i^{(l)}\Big), & \text{output layer} \\ \Big\|_{k=1}^{K}\sigma\Big(\sum_{j\in\mathbb{N}_i}\alpha_{ij,k}^{(l)} W_k^{(l)} h_i^{(l)}\Big), & \text{else} \end{cases} \tag{9}$$

where $\alpha_{ij,k}^{(l)}$ and $W_k^{(l)}$ represent the normalized attention coefficients and weight matrix in the k-th head, respectively. In addition, the number of heads is set to K.

In Section 2.1, we propose a road network transformation method to transform the urban road network into a weighted directed graph. It is noted that the edge features are ignored in the GAT layer mentioned above. Hence, to take advantage of the edge features, we introduce the edge weights E of the constructed weighted directed graph as prior knowledge to improve the process of calculating the attention weights. Furthermore, introducing the edge features into the GAT layer can also reduce the impacts of the useless connections caused by construction errors:

$$\alpha_{ij} = \frac{\exp(E_{ij}\cdot e_{ij})}{\sum_{k\in\mathbb{N}_i}\exp(E_{ik}\cdot e_{ik})} \tag{10}$$

3.3 | Temporal dependencies

In Equation (4), the lower-level features are mapped into a high-dimensional space through a linear transformation. It is noted that the input features of each node are traffic states of the previous s time steps so that they can be regarded as time series. As fully capturing the time-varying characteristics of the traffic flow is the key to accurately predicting future traffic states, this linear transformation strategy is unable to adequately explore the temporal evolution of the input features. In the field of time series processing, RNNs and their widely used variants, LSTM and GRU, have been applied in existing studies. Because GRU is simple to compute and easily converge (Cho et al., 2014; Gu et al., 2019; Zhao et al., 2019), we utilize them to transform the input features into higher-level features instead of using a linear transformation, aiming at capturing the temporal dependencies of traffic flows effectively.

Specifically, if the input features of the graph are denoted as $F \in R^{N\times r}$, where N is the number of nodes, and r represents the number of previous time steps, the output of Equation (4) can be written as $F' \in R^{N\times r'}$, $r' > r$. The linear transformation utilizes a learnable weight matrix $W \in R^{r\times r'}$ to map the input features into r'-dimensional space. When GRU is used, the input features are reshaped as $F \in R^{N\times r\times 1}$ first. Assuming that the number of hidden units in the GRU is set as g, the dimension of the output is $R^{N\times r\times g}$, it can be reshaped as a 2D tensor $R^{N\times(r\times g)}$, which can also achieve higher-level feature transformation.

There are two components in a GRU block: A reset gate r_t and an update gate z_t. The former is utilized to discard historical information that is unrelated to future states, and the latter can help to capture long-term dependencies in time series. Given an input feature q_t, the GRU layer can be expressed as follows:

$$r_t = \sigma(W_{rq}q_t + W_{rh}H_{t-1} + b_r) \tag{11}$$

$$z_t = \sigma(W_{zq}q_t + W_{zh}H_{t-1} + b_z) \tag{12}$$

$$\tilde{H}_t = \phi(W_{hq}q_t + W_{hh}(r_t \odot H_{t-1}) + b_h) \tag{13}$$

$$H_t = z_t \odot H_{t-1} + (1 - z_t)\odot \tilde{H}_t \tag{14}$$

$$\sigma(x) = \frac{1}{1+\exp(-x)} \tag{15}$$

$$\phi(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)} \tag{16}$$

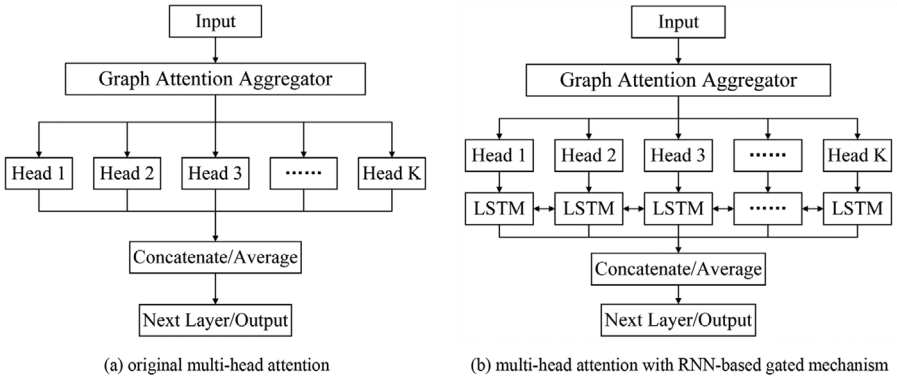


FIGURE 4 Multi-head attention with recurrent neural network-based gated mechanism

where H_t denotes the hidden states at time t, and W, b are the weight matrix and bias, respectively.

3.4 | Gated mechanism

Since the multi-head attention mechanism can aggregate information from multiple representation subspaces (Vaswani et al., 2017), the training process can be highly stable. However, it regards the importance of all subspaces as equal, ignoring the differences between different heads. Moreover, when the multi-head attention mechanism is applied on graphs, the representation subspaces of specific nodes may not even exist (J. Zhang, Shi, et al., 2018). As illustrated in Figure 4(a), in the multi-head attention mechanism, the input features are fed into K different heads, and K outputs are obtained. J. Zhang, Shi et al. (2018) employed a CNN-based gated aggregator to determine the importance of each node at different heads for graph learning. However, the proposed CNN-based gated aggregator only relies on neighboring information, ignoring the interactions in the multi-head mechanism. In this study, we treat the outputs of different attention heads as sequence data and propose an RNN-based soft gated mechanism, which is shown in Figure 4(b), to determine the importance of each head. In this way, the outputs of each head through the gated mechanism are not only determined by themselves but are also impacted by other heads.

As there is no specific direction for the information dissemination process among all the attention heads, we employ the BiLSTM to reflect the interactions within them. Different from LSTM, the BiLSTM adds a hidden layer that allows information to pass from back to front, enhancing its ability to handle backpropagated information. Additionally, after passing through the gated layer, all the outputs of the BiLSTM are fed into Equation (9). Here, the output dimension of the BiLSTM layer is the same as that of its input.

Supposing that g_t denotes the output of head t, the details of the naive LSTM layer are written below:

$$I_t = \sigma(W_{ig} g_t + W_{ih} H_{t-1} + b_i) \tag{17}$$

$$F_t = \sigma(W_{fg} g_t + W_{fh} H_{t-1} + b_f) \tag{18}$$

$$O_t = \sigma(W_{og} g_t + W_{oh} H_{t-1} + b_o) \tag{19}$$

$$\tilde{C}_t = \phi(W_{cg} g_t + W_{ch} H_{t-1} + b_c) \tag{20}$$

$$C_t = F_t \odot C_{t-1} + I_t \odot \tilde{C}_t \tag{21}$$

$$H_t = O_t \odot \phi(C_t) \tag{22}$$

Here, I_t, F_t, O_t, C_t, and H_t denote the input gate, forget gate, output gate, memory cell, and hidden state, respectively. Based on this directional LSTM, the BiLSTM achieves the bidirectional propagation of information by calculating $\vec{H}_t$ and $\overleftarrow{H}_t$, where "$\rightarrow$" and "$\leftarrow$" denote the forward and backward directions, respectively. After obtaining these two hidden states, the hidden state H_t can be generated by concatenating $\vec{H}_t$ and $\overleftarrow{H}_t$.

3.5 | STGGAT

In this study, a deep learning framework named STGGAT is proposed for traffic flow prediction on urban road networks. To extract the spatiotemporal dependencies from

the constructed weighted directed graphs, STGGAT integrates a GRU layer, a GAT layer, and an RNN-based gated mechanism. We also adopt a residual structure (He et al., 2016) to accelerate the convergence of the proposed STGGAT model.

Suppose that $\psi(x)$ denotes the ideal mapping for predicting the target values y based on the input features x. The aim of the residual structure is to fit the residual mapping $\psi(x) - x$. The residual structure opens a highway that allows the input to be directly propagated to the output. Its core innovation is allowing the information to span several layers, thereby improving the forward speed of information. Therefore, in practical applications, residual mapping is easy to optimize and captures the subtle fluctuations of identity mapping.

After fully expanding all terms, the proposed STGGAT model with a residual structure can be formulated as follows:

$$\mathrm{TF}_k^{(l)}(x) = \begin{cases} \mathrm{GRU}_k^{(l)}(x), & l = 1 \\ W_k^{(l)} x, & l > 1 \end{cases} \tag{23}$$

$$\alpha_{ij,k}^{(l)} = \frac{\exp(E_{ij} \cdot \mathrm{LeakyReLU}(\vec{a}^{\mathrm{T}}[\mathrm{TF}_k^{(l)}(\mathrm{h}_i^{(l)}) || \mathrm{TF}_k^{(l)}(\mathrm{h}_j^{(l)})]))}{\sum_{p \in \mathbb{N}_i} \exp(E_{ip} \cdot \mathrm{LeakyReLU}(\vec{a}^{\mathrm{T}}[\mathrm{TF}_k^{(l)}(\mathrm{h}_i^{(l)}) || \mathrm{TF}_k^{(l)}(\mathrm{h}_p^{(l)})]))} \tag{24}$$

$$SF_{i,k}^{(l)} = \sum_{j \in \mathbb{N}_i} \alpha_{ij,k}^{(l)} \cdot \mathrm{TF}_k^{(l)}(h_j^{(l)}) \tag{25}$$

$$\widetilde{SF}_i^{(l)} = \{\widetilde{SF}_{i,1}^{(l)}, \ldots, \widetilde{SF}_{i,K}^{(l)}\} = \mathrm{BiLSTM}(SF_{i,1}^{(l)}, \ldots, SF_{i,K}^{(l)}) \tag{26}$$

$$h_i^{(l+1)} = \begin{cases} \sigma\left(\frac{1}{K}\sum_{k=1}^{K} \widetilde{SF}_{i,k}^{(l)}\right) + W_{res}^{(l)} h_i^{(l)}, & \text{output layer} \\ \Big\|_{k=1}^{K} \sigma(\widetilde{SF}_{i,k}^{(l)}) + W_{res}^{(l)} h_i^{(l)}, & \text{else} \end{cases} \tag{27}$$

Here, $W_{res}^{(l)}$ is a learnable weight matrix used to make the dimension of input features equal to that of the outputs.

In summary, compared with the naive GAT model, the proposed STGGAT model expresses a more powerful capability to extract spatiotemporal dependencies. For the temporal correlations, a GRU layer is employed to mine the dynamic variation characteristics among the previous traffic states. For spatial relationship modeling, edge features are introduced into the self-attention mechanism as prior knowledge. Moreover, an RNN-based gated module is proposed to determine the different importance within the multi-head attention mechanism.

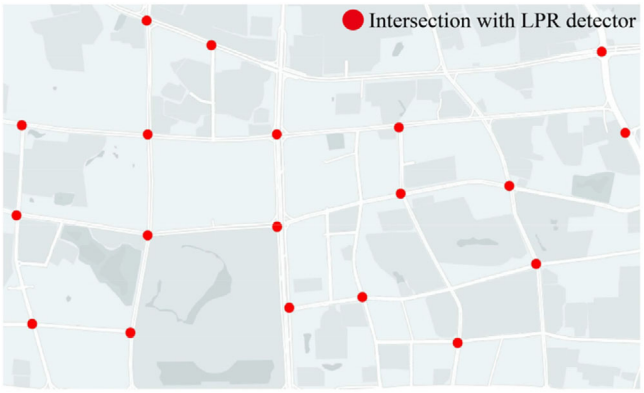


FIGURE 5 An overview of the study area and the coverage of license plate recognition devices

4 | DATA DESCRIPTION

The dataset used in this study was collected by the LPR system in Changsha, China. Based on video processing technologies, LPR devices can obtain collection times, car numbers, directions, lane numbers, and so forth. Due to their ability to obtain full samples at intersections and extract traffic states such as average speed, travel time, vehicle trajectory, and so forth, LPR devices have been widely used in travel behavior analysis (H. Chen et al., 2017) and origin-destination pattern estimation (Rao et al., 2018). From the LPR records, three scales of traffic states can be obtained: Lane-level, approach-level, and intersection-level. It is obvious that the micro traffic volumes express more fluctuations and variations, while the macro volumes are more stable. In this paper, traffic volumes in a local road network in southern Changsha, illustrated in Figure 5, are selected for the experiment in this case study. This area contains 19 intersections, 64 approaches, and 301 lanes, including arterial roads, secondary trunk roads, and branches. Considering their potential assistance in the applications of traffic guidance and control, we select traffic volumes at the lane level and approach level for analysis. From the monthly average daily traffic (MADT) distributions, illustrated in Figure 6, traffic flows on urban roads are generally not very heavy, and approach-level traffic volumes are much higher than those at the lane level. The selected LPR dataset was collected from July 1 to 31, 2019, and consists of 19,441,883 records in total. All license plate information is masked to protect drivers' privacy. We aggregate the 5-min traffic volumes at the lane and approach levels and impute the missing data using historical average volumes before performing predictions. To evaluate the performance of the proposed model, we divide the dataset into a training set, validation set, and test set at a ratio of 60%: 20%: 20%. Here, the training and validation sets are employed for model

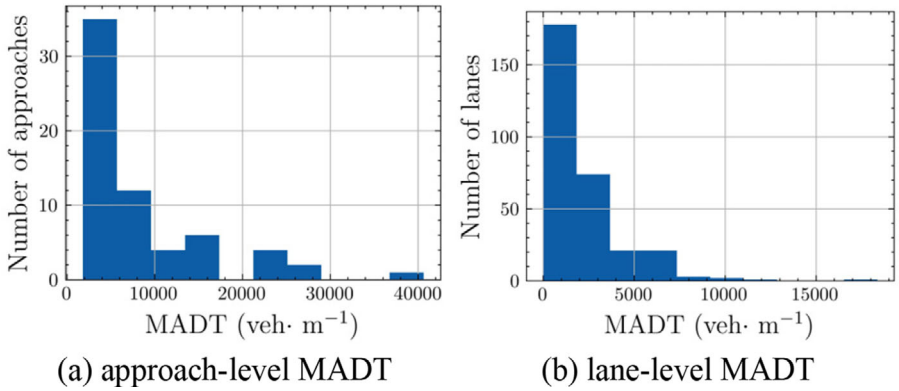


FIGURE 6 Monthly average daily traffic distribution in the selected area

training and hyperparameter tuning, and the test set is utilized for the final model performance testing.

5 | EXPERIMENT

5.1 | Performance metrics

To evaluate the performance of the prediction model, we employ three indicators to measure the error between the ground truths and predictions, namely, root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE):

$$\text{RMSE} = \sqrt{\frac{1}{n \cdot N} \sum_{i=1}^{n} \sum_{j=1}^{N} \left(\tilde{y}_j^i - y_j^i\right)^2} \tag{28}$$

$$\text{MAE} = \frac{1}{n \cdot N} \sum_{i=1}^{n} \sum_{j=1}^{N} \left|\tilde{y}_j^i - y_j^i\right| \tag{29}$$

$$\text{MAPE} = \frac{1}{n \cdot N} \sum_{i=1}^{n} \sum_{j=1}^{N} \left|\frac{\tilde{y}_i^j - y_i^j}{y_i^j}\right| \times 100\% \tag{30}$$

In the above equations, y_i^j, $\tilde{y}_i^j$, N and n represent the ground truths, predictions, number of detectors, and number of test samples, respectively. Specifically, the smaller the values of these three indicators are, the higher the prediction accuracy.

5.2 | Determination of the model parameters

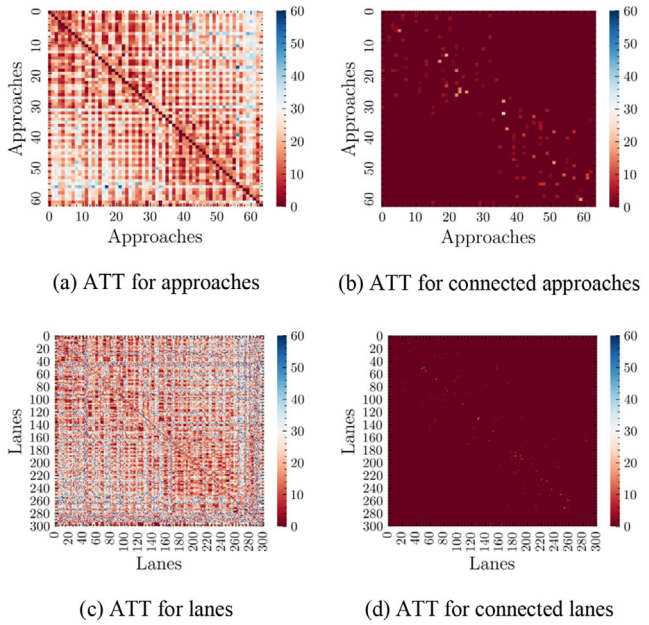
For the temporal dimension of the input data, the number of previous time steps s is set as 10. Adam (Kingma & Ba, 2014) is adopted as the optimizer with a learning rate of 10^{-3}, and we employ the mean square error as the loss function. Limited by the memory of the GPU, we set the batch size of the approach-level prediction as 128 and that of the lane-level prediction as 16. In the weighted directed graph construction method, the hyperparameters k and λ have an essential impact on the sparseness of the graph and the node degrees. In detail, if the graph is too dense, it may introduce several worthless edges; otherwise, if it is too sparse, some critical information may not be delivered. Furthermore, the hidden dimension of the GRU, l, also plays a vital role in determining the prediction accuracy. Hence, to select the best hyperparameters for the proposed prediction framework, we carefully apply a grid search strategy on $k \in \{1, 3, 5, 7, 10\}$ and $l \in \{16, 20, 24, 28, 32\}$. For λ, we follow (Cupertino et al., 2013) and set it as 3. After the grid search process is completed, $k = 3$, $l = 28$ are selected as the best hyperparameters. Based on the hyperparameters obtained for these settings, the properties of the constructed graph are shown in Table 1. In addition, Figure 7 illustrates the ATT matrices of different sections in the road network. The ATT matrices shown in Figures 7(b) and (d) only represent the ATT between connected sections, and they can be calculated by $ATT \odot A$, where ATT and A denote the ATT and adjacency matrices, respectively. It is found that the road network transformation method plays the role of a travel time filter, which only allows nodes that are "proximal in time" to connect.

5.3 | Performance comparison

In this subsection, the proposed STGGAT model is compared with other widely used traffic flow prediction models, including statistical methods, machine learning methods, and advanced deep learning methods. The following is a brief introduction to these baseline models, where the

TABLE 1 The properties of the constructed graph

	Approach-level	Lane-level
Node number	64	301
Edge number	208	1053
Average degree	6.50	7.00
Density	0.05	0.01
Average clustering	0.20	0.09



(a) ATT for approaches (b) ATT for connected approaches

(c) ATT for lanes (d) ATT for connected lanes

FIGURE 7 Average travel time between different road sections

grid search strategy searches the hyperparameters of all the models.

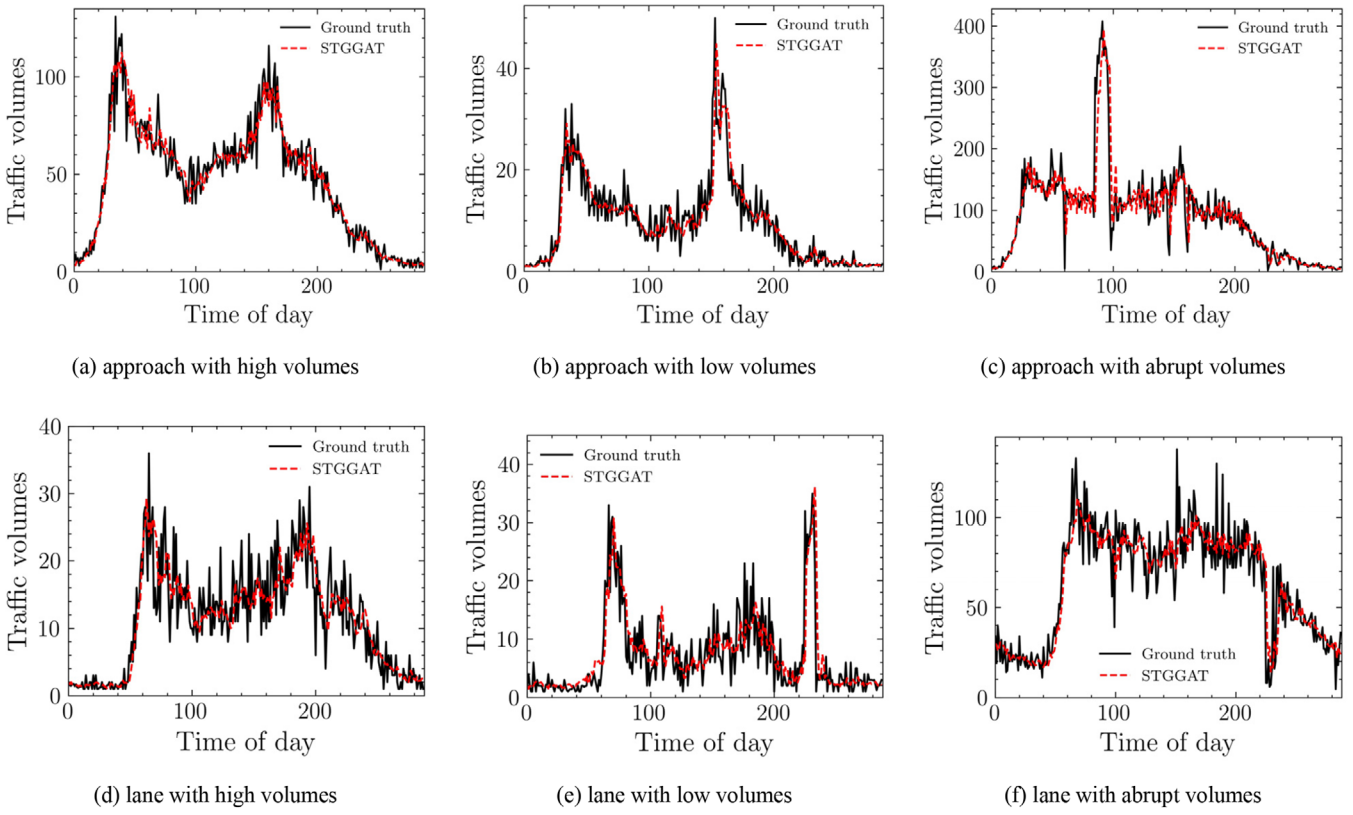
1. **ARIMA**: Autoregressive integrated moving average (ARIMA) is one of the most widely used statistical models in traffic flow prediction. The parameters of ARIMA are estimated from the maximum likelihood estimation by the Statsmodels package (Seabold & Perktold, 2010) in Python. In addition, the hyperparameters of the ARIMA model are set as $p = 5$, $d = 0$, $q = 4$ for both tasks.
2. **BPNN**: A three-layer backpropagation neural network (BPNN) is established, and it consists of an input layer, a hidden layer with ReLU as the activation function, and an output layer. Specifically, the number of hidden units in the hidden layer is set to 64 for approach-level prediction and 240 for lane-level prediction.
3. **LSTM**: This is the LSTM network (Ma et al., 2015) described in Section 3.4. The numbers of hidden units of LSTM are set to 320 for both prediction tasks.
4. **GCN**: This is the naive GCN (Kipf & Welling, 2016). The number of hidden dimensions is set to 240 for approach-level prediction and 200 for lane-level prediction.
5. **GAT**: This is the naive GAT (Veličković et al., 2017) described in Section 3.2. The numbers of hidden dimensions are set to 120 and 240.
6. **LSGC-LSTM**: This is a two-layer stacking model combining the localized spectral graph convolution neural network (LSGC; Defferrard et al., 2016) with an LSTM layer. In this study, the hop of the graph convolution operation is set to 3, and the number of hidden units in the hidden layer is set to be equal to the number of nodes, N.
7. **T-GCN**: The temporal GCN (T-GCN), proposed by Zhao et al. (2019), combines the GCN with GRU. The T-GCN is implemented based on the source code on GitHub.[2] The numbers of hidden units are set to 240 and 200 for these two prediction tasks.
8. **TGC-LSTM**: This is the traffic graph convolutional RNN (TGC-LSTM) proposed by Cui et al. (2019). We utilize the source code[3] shared by the authors to construct this prediction model. We utilize the travel time matrix to calculate the free-flow reachable matrix instead of the free-flow speed. The number of hidden units in the hidden layer is set to be equal to the number of nodes, N.

Among the baselines, all the GNN-based models are applied on the directed graph without edge weights based on the road network transformation method proposed in Section 2.1.2. All the experiments are conducted on Python 3.6.10 with a Windows workstation (RAM: 32GB, CPU: Intel Core (TM) i9-9900K @ 3.6 GHz, GPU: NVIDIA 2080Ti). The proposed STGGAT model is implemented utilizing the open-source GNN framework Deep Graph Library (DGL; Wang et al., 2019) with MXNet (Chen et al., 2015) as the backend. In addition, all the deep learning methods are trained with an early stop strategy to avoid overfitting. Table 2 shows the prediction performances of the aforementioned models for both approach- and lane-

[2] https://github.com/lehaifeng/T-GCN.
[3] https://github.com/zhiyongc/Graph_Convolutional_LSTM.

TABLE 2 Prediction performance of different models for both approach-level and lane-level prediction

Model	Approach-level			Lane-level		
	RMSE	MAE	MAPE (%)	RMSE	MAE	MAPE (%)
ARIMA	8.320	4.995	39.896	2.891	1.803	43.535
BPNN	8.436	5.044	39.536	2.918	1.814	43.254
LSTM	8.553	4.974	34.726	2.902	1.775	41.392
GCN	13.533	6.261	37.334	3.883	1.991	44.145
GAT	8.164	4.868	35.782	2.885	1.779	40.405
LSGC-LSTM	10.251	5.340	39.915	2.941	1.806	42.707
T-GCN	11.612	8.363	129.262	3.593	2.446	76.596
TGC-LSTM	8.547	4.867	34.812	3.005	1.863	43.566
STGGAT	**7.732**	**4.546**	**31.800**	**2.828**	**1.742**	**39.462**



(a) approach with high volumes (b) approach with low volumes (c) approach with abrupt volumes

(d) lane with high volumes (e) lane with low volumes (f) lane with abrupt volumes

FIGURE 8 Comparison of ground truths and predicted results on July 31, 2019

level predictions, where the indicators with the best performances are marked in bold. Prediction comparisons of six selected case studies involving different traffic conditions, including high, low, and abrupt volumes, are displayed in Figure 8. From the overall prediction results, several critical conclusions can be drawn below:

1. The proposed STGGAT model expresses superior accuracy to those of baselines due to its powerful ability to perform spatiotemporal dependency extraction. Specifically, the temporal evolution can be captured by the GRU layer, and the spatial correlations can be mined by the GAT aggregator and prior edge features. For lane-level prediction, as the total traffic volumes of each lane at a 5-min scale are generally low (shown in Figure 6(b)), the prediction errors of all the models are relatively low and similar. It is also noted that the naive GCN and its variants, LSGC-LSTM and T-GCN, have much lower accuracies than those of other models. The reason for this may be caused by the poor ability of these models to capture intense fluctuations in traffic flows on urban roads. In terms of

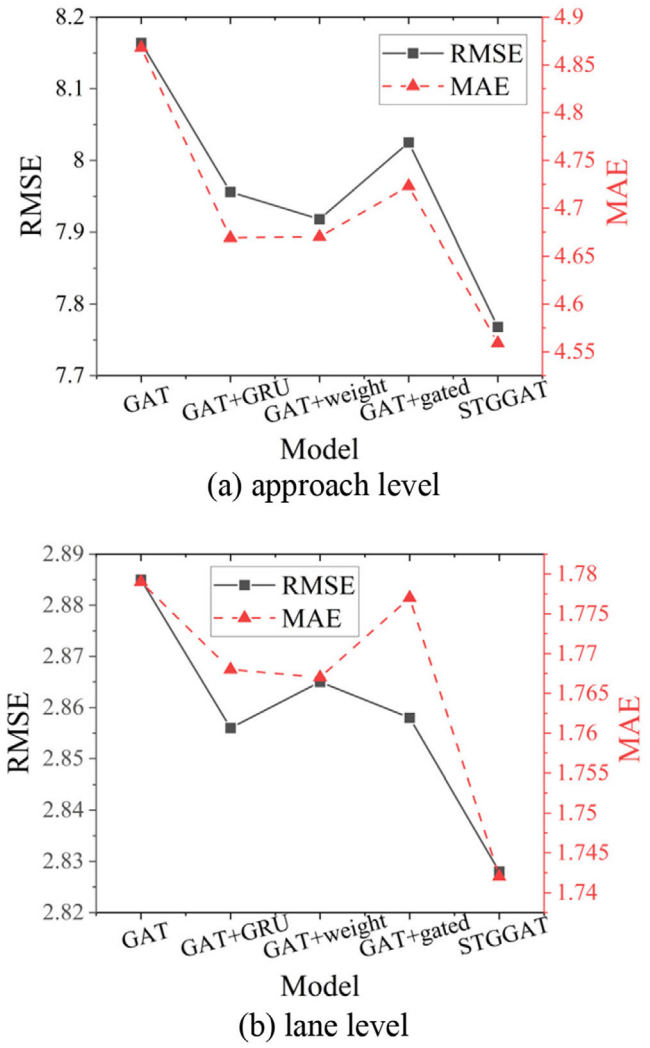


FIGURE 9 Component analysis of spatiotemporal gated graph attention network (STGGAT)

the TGC-LSTM, as the free-flow speed matrix plays a vital role in filtering the useless spatial information and recognizing the influential road segments, it achieves a relatively accurate performance among the GCN models.

2. To explore the effects of the components in STGGAT, the prediction performance of each component is illustrated in Figure 9. According to Figure 9, all the components utilized in STGGAT can improve the prediction accuracy of the naive GAT. Among all the components, since it introduces the volume transition relationships as prior knowledge, the GAT with edge weights obtains the largest improvement over the naive GAT. This may be because there exists a robust, strong spatial correlation in the urban road network that even outweighs the temporal evolution, and it can be extracted by the weighted directed network construction method developed in this study. Furthermore, the GAT with GRU can capture the temporal evolution characteristics in the historical traffic data, and the proposed RNN-based gated module is effective in distinguishing the importance of each head in the multi-head mechanism. In total, these components can explore the characteristics of traffic flows from different perspectives, leading to an accurate and stable prediction performance for STGGAT.
3. The multistep prediction results are illustrated in Figure 10. From Figure 10, we can see that for all the models, the prediction errors increase as the forecast time step increases, and STGGAT consistently achieves relatively high and stable performances. As shown in Figure 10(a), for the 10-step prediction of approach-level traffic volumes, the TGC-LSTM achieves higher accuracies than those of STGGAT. This may be because the TGC-LSTM utilizes a k-hop sampling strategy, so not only the information pertaining to the adjacent roads but also that of k-hop roads can be fed to the prediction model, while the GAT-based model only uses the first neighbor. Furthermore, as the spatial dependencies in the road network play an important role in accurate traffic flow prediction, models considering spatiotemporal correlations express more stable performances than those of models that only include temporal evolution. Although ARIMA and BPNN perform well in the one-step prediction of fluctuating traffic conditions, the prediction performances of these methods are not stable for long prediction steps. For ARIMA, the subsequent predictions are determined by the prior predicted results, so the prediction error of the multistep model accumulates continuously.
4. Although the performance indicators in Table 2 show the prediction performances from a network-scale perspective, it is also necessary to further evaluate the prediction error of each node. Figure 11 illustrates the RMSE and MAPE distributions of STGGAT in terms of prediction at the approach and lane levels. Overall, the prediction error distributions of these two tasks show similar trends: The RMSE is concentrated at a low level, while the MAPE presents an approximately symmetrical distribution. Specifically, the RMSEs of approximately 90% of the approaches and lanes are lower than 15 and 6, respectively. For the MAPE, the symmetry axis of MAPE is 32% for approach-level and 45% for lane-level prediction tasks. Moreover, it is noted that the distribution of RMSE is consistent with the MADT distribution in Figure 6. This reveals that the prediction RMSE is profoundly affected by traffic volume magnitude. Figure 12 illustrates the boxplot of RMSE and MAE distributions of all the approaches and lanes. For the approach-level prediction, the abnormal values of the proposed STGGAT model, especially the maximum

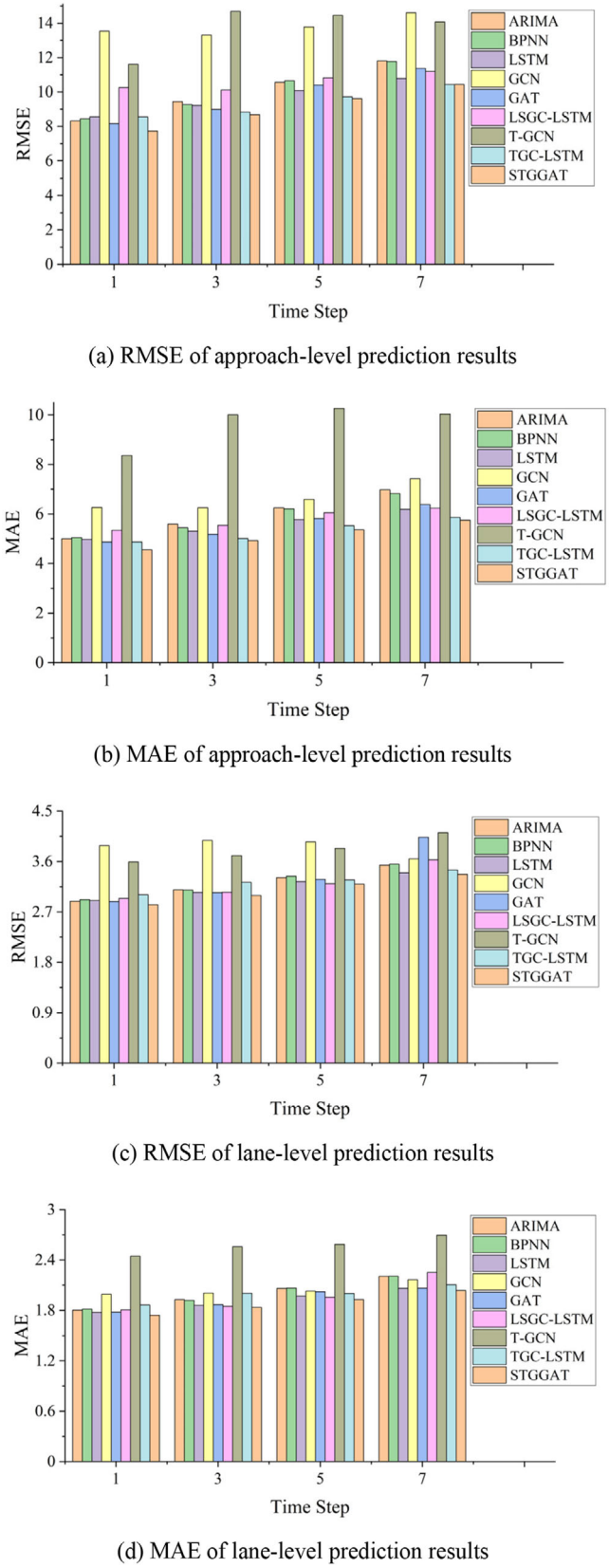


FIGURE 10 Comparison of multi-step prediction performance

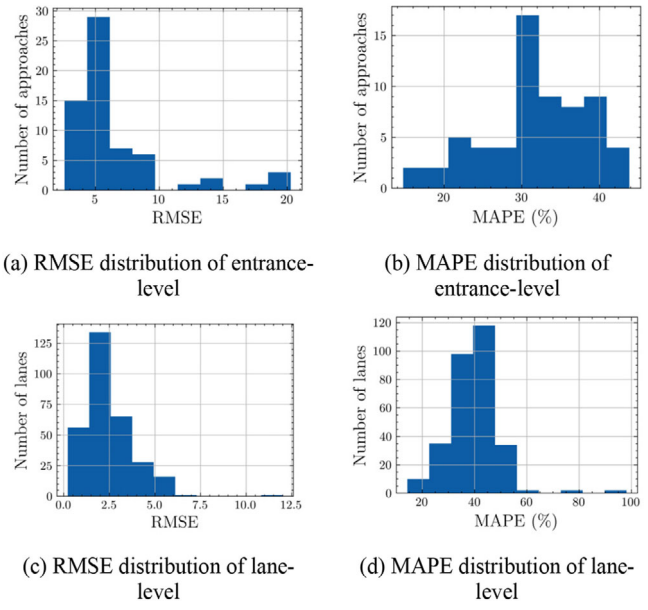


FIGURE 11 Prediction error distributions of STGGAT

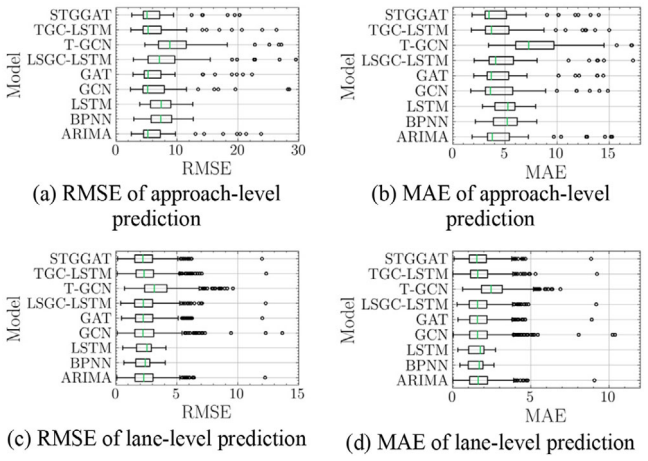


FIGURE 12 Boxplot of prediction error distribution under approach and lane levels

error, are much lower than those of all the other models. Thus, STGGAT exhibits a superior prediction capability to those of other models in intense situations and for high volumes. Additionally, based on the lower and upper quartile values, STGGAT also outperforms other models, and this further proves the effectiveness of the proposed model. From Figures 12(c) and (d), STGGAT has fewer abnormal data points than other models, and these outliers are closer to normal values. In summary, although the approach- and lane-level traffic volumes exhibit strong fluctuations, the proposed STGGAT model can improve upon the prediction performances of baseline models by mining the spatiotemporal dependencies of the urban road network.

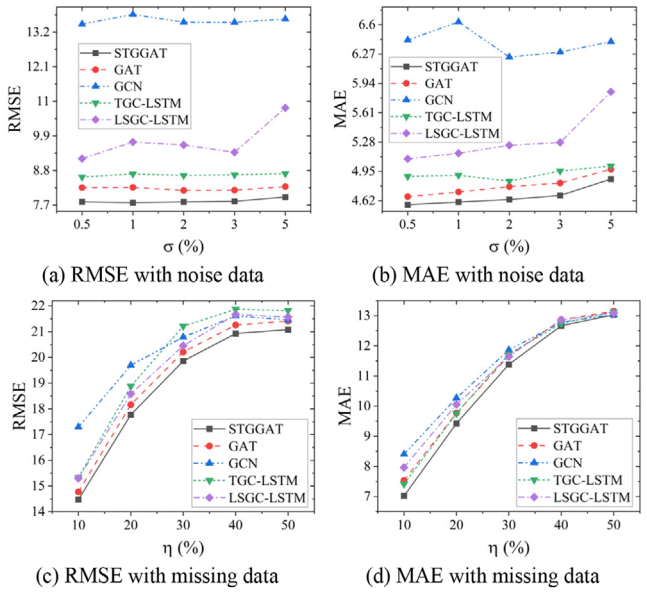


FIGURE 13 Perturbation and robustness experiments for approach-level prediction

5.4 | Fault tolerance analysis

In the application of a data-driven traffic flow prediction model, it is inevitable that we will encounter the missing data problem, which usually leads to incorrect predictions and responses (X. Chen et al., 2019). To further explore the robustness of the proposed STGGAT model to perturbations, we insert random noise and stochastic missing data into the LPR dataset and validate the fault tolerance of STGGAT. Specifically, the Gaussian distribution $P \in (0,\ \sigma^2), \sigma \in \{0.5\%,\ 1\%,\ 2\%,\ 3\%,\ 5\%\}$ of the average volumes and the missing rates $\eta \in \{10\%,\ 20\%,\ 30\%,\ 40\%,\ 50\%\}$ are utilized to generate the dataset for this robustness analysis.

The robustness analysis for approach-level traffic prediction with regard to perturbations is illustrated in Figure 13. Under noise situations, all GNN models achieve relatively stable performances, and STGGAT always outperforms the other methods, demonstrating that these models are less impacted by noise data than the other models. In terms of missing data situations, the prediction accuracies of all the models decline gradually as the rate of missing data increases. Similarly, STGGAT obtains the best prediction results, indicating that it is more fault-tolerant than other GNN models. However, although STGGAT is superior, missing data are still a critical factor that has a significant impact on forecasting accuracy. In practical applications, it is necessary to perform effective missing data imputation before performing predictions.

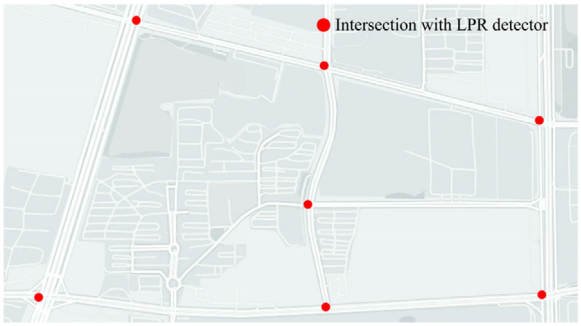


FIGURE 14 The selected local road network to compare the prediction performance of the constructed graph and the physical network

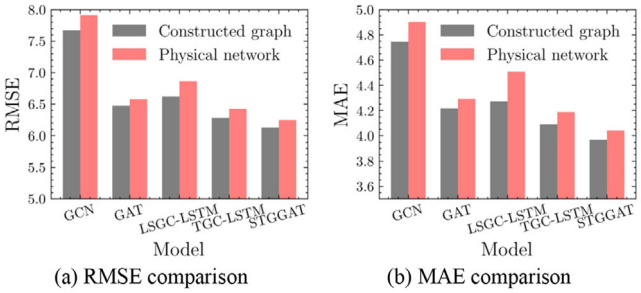


FIGURE 15 Prediction performance on the constructed graph and the physical road network

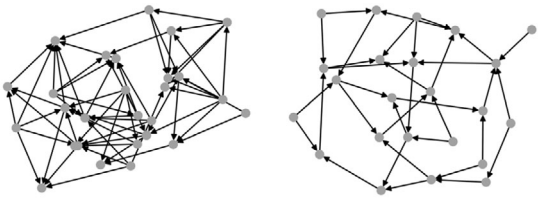
5.5 | Road network transformation analysis

This study develops a road network transformation method to construct weighted directed graphs for the purpose of enhancing spatial dependencies. It is necessary to conduct a comparison experiment on the constructed graph and the physical network to evaluate whether this method is useful for traffic flow prediction. However, the road network shown in Figure 5 is too sparse to construct a topology graph based on the adjacent relationships. Thus, we select a local road network in northern Changsha, illustrated in Figure 14, to establish the adjacency-based topology graph for approach-level traffic flow prediction. To distinguish between these two graphs, we call them the constructed graph and the physical network.

Although LPR devices in this area are densely equipped, there still exist several intersections without detectors. When establishing the adjacency-based graph, we ignore the turning options, so only the driving decision to go straight is considered. Figure 15 shows the comparison of all the GNN-based model predictions on these two graphs. We can see that the constructed graph always outperforms the physical network in terms of prediction accuracy, and the proposed STGGAT model consistently achieves

TABLE 3 The properties of the constructed graph and the physical road network

	Constructed graph	Physical network
Node number	27	27
Edge number	88	70
Average degree	6.52	5.19
Density	0.13	0.10
Average clustering	0.24	0.00



(a) constructed graph using the proposed method (b) physical network

FIGURE 16 The topology structure of the constructed graph and the physical network

the best performance. Furthermore, from the properties and topological structures of these two graphs (shown in Table 3 and Figure 16), we can find that the constructed graph is denser than the physical network. More interactions exist among the neighbors according to the values of average clustering. This interesting finding demonstrates the constructed graph's superior ability to that of the physical network with regard to extracting the spatial dependencies of the urban road network. The reason for this may be that the constructed graph has a more substantial capability than that of the physical network for spatial dependency mining by connecting correlated nodes.

5.6 | Inductive learning

Compared with the GCN, a major advantage of the GAT lies in its ability to perform inductive learning tasks. It can achieve accurate prediction results on graphs whose structures do not appear in the training set. In the traffic flow prediction task, it can be said that a GAT model pretrained on a specific road network can be applied on another road network. This is an extremely difficult task for a GCN but one that is useful in real-world applications. In this subsection, to explore the inductive learning ability of the proposed STGGAT model, we employ the pretrained STGGAT on the road network shown in Figure 5 to predict the traffic volumes at the approach level for the road network illustrated in Figure 14.

The prediction results of different models are shown in Table 4, and we denote the inductive learning predictions of the GAT and STGGAT as GAT-i and STGGAT-i, respectively. Different from the calculation operations on the Laplace matrices of graphs such as the GCN, the weights in the GAT only rely on the node features. Therefore, the GAT can be applied to graphs with different structures without undergoing training again. From the inductive learning prediction performance, it can be seen that the GAT-based models achieve relatively acceptable accuracy compared with those of the baselines. In addition, the STGGAT-i model also outperforms the GAT-i model. This may be because the STGGAT model introduces edge features into the prediction model as prior knowledge, so the actual traffic characteristics of the newly seen road network are involved in the prediction model. Thus, STGGAT may reduce the dependence of prediction results on model parameters, leading to a stronger generalization ability than those of other models.

6 | CONCLUSION

Due to the presence of complicated spatiotemporal dependencies, the accurate prediction of network-scale traffic volumes at intersections is a significant challenge. This paper presents a forecasting framework, the STGGAT, to fill this gap and predict approach- and lane-level traffic volumes at urban intersections. A complex network construction method is employed to transform an urban road network to a weighted directed graph based on the ATTs and volume transition relationships between different approaches or lanes. The proposed STGGAT integrates four essential components: A GRU layer, a GAT, an RNN-based gated mechanism, and a residual structure. In detail, GRU is introduced to transform the input features into high-level features and capture the temporal evolution of the traffic volumes. The edge weights extracted from the volume transition relationships are aggregated to the attention coefficients as prior knowledge, with the aim of enhancing the capability of GAT to deal with edge information. Moreover, a BiLSTMlayer is adopted to determine the

TABLE 4 Prediction performance of different models on the local road network shown in Figure 14

Model	RMSE	MAE	MAPE (%)
ARIMA	6.518	4.260	37.663
BPNN	6.560	4.279	37.664
LSTM	6.585	4.271	35.972
GCN	7.670	4.745	42.217
GAT	6.473	4.217	39.569
LSGC-LSTM	6.622	4.272	38.497
TGC-LSTM	6.283	4.090	37.216
STGGAT	6.129	3.970	35.347
GAT-i	6.669	4.312	40.152
STGGAT-i	6.535	4.163	36.300

Note: GAT-I and STGGAT-I are inductive learning predictions of the GAT and STGGAT.

importance of different heads in the multi-head attention mechanism. The residual structure is utilized to accelerate the convergence process.

Validated on the LPR system in Changsha, China, the proposed STGGAT is compared with several baselines, including statistical models, machine learning methods, and advanced graph neural networks. The experimental results demonstrate the superior accuracy of STGGAT in terms of both approach- and lane-level prediction tasks and its stability in multistep prediction. Through a fault tolerance analysis regarding noise and missing data, the robustness of STGGAT is also demonstrated. Furthermore, by validating it on a small road network in a different structure, we find that the STGGAT model can achieve acceptable performance on an inductive learning task and prove the graph construction method's superior effectiveness to that of the physical road network. From the overview of this study, several interesting conclusions can be summarized.

1. The intermittent interruption effect of signal control at urban intersections leads to approach- and lane-level traffic volumes exhibiting extreme fluctuations. In this case, traditional traffic flow prediction methods may encounter bottlenecks, especially in multistep prediction, while the models considering spatiotemporal dependencies achieve stable and robust prediction performances.
2. Although many deep learning models have been proposed to address traffic prediction, we need to explore traffic characteristics to make the models optimally suitable for traffic problems. For instance, there are several differences in the spatial dependencies between adjacent roads under the combined effects of multiple factors. This is a critical issue, but it is overlooked in many studies. To fill this gap, this study proposes a road network transformation method to construct a weighted directed graph based on the extracted ATTs and volume transition relationships. This method allows nodes with low travel times to be connected and determines each edge's importance by its volume transition relationships. In this way, the constructed graph is highly suitable for solving traffic volume prediction issues. According to the experimental results, this method can improve the network-scale traffic prediction performances of existing models.
3. Inductive learning is another concern in traffic flow prediction research. A prediction model with inductive learning ability could not only reduce training costs significantly but also save storage space. Although the proposed STGGAT model achieves good prediction performances on inductive learning tasks, it still needs further improvements.

There are several potential extensions for future studies in this field. For instance, traffic volumes are affected by multiple factors, such as traffic accidents and weather. In this study, only historical traffic volumes and spatial dependencies are adopted in the prediction model. Moreover, the high-frequency fluctuations at the short scale make achieving accurate traffic volume predictions at urban intersections a significant challenge. It would be an interesting topic for future works to improve the prediction performance by introducing several denoising methods, such as wavelet theory (Adeli & Karim, 2005) and ensemble EMD (Zhang et al., 2020). Besides, many researchers also explored the characteristics of traffic flow time series from their complex network structures (Yan et al., 2017). Exploring the complex network structure of traffic systems from the temporal and spatial dimensions may also improve the prediction accuracy. Furthermore, a large number of novel and powerful algorithms have emerged with the flourishment of artificial intelligence, such as enhanced probabilistic neural network (Ahmad-

lou & Adeli, 2010), neural dynamic classification algorithm (Rafiei & Adeli, 2017), dynamic ensemble learning algorithm (Alam et al., 2020), and finite element machine (Pereira et al., 2020). These emerging models may open a new chapter in the field of traffic flow prediction.

REFERENCES

Adeli, H., & Ghosh-Dastidar, S. (2004). Mesoscopic-wavelet freeway work zone flow and congestion feature extraction model. *Journal of Transportation Engineering*, *130*(1), 94–103.

Ahmadlou, M., & Adeli, H. (2010). Enhanced probabilistic neural network with local decision circles: A robust classifier. *Integrated Computer-Aided Engineering*, *17*(3), 197–210.

Alam, K. M. R., Siddique, N., & Adeli, H. (2020). A dynamic ensemble learning algorithm for neural networks. *Neural Computing and Applications*, *32*(12), 8675–8690.

Bao, L., Ma, B., Chang, H., & Chen, X. (2019). Masked graph attention network for person re-identification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Long Beach, CA, pp. 1496–1505.

Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, *15*(6), 1373–1396.

Chen, H., Yang, C., & Xu, X. (2017). Clustering vehicle temporal and spatial travel behavior using license plate recognition data. *Journal of Advanced Transportation*, *2017*(7), 1–14.

Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., & Zhang, Z. (2015). MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems. http://arxiv.org/abs/1512.01274

Chen, X., He, Z., Chen, Y., Lu, Y., & Wang, J. (2019). Missing traffic data imputation and pattern discovery with a Bayesian augmented tensor factorization model. *Transportation Research Part C: Emerging Technologies*, *104*, 66–77.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *EMNLP 2014–2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1724–1734.

Cui, Z., Henrickson, K., Ke, R., & Wang, Y. (2019). Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting. *IEEE Transactions on Intelligent Transportation Systems*, *21*(11), 4883–4894.

Cui, Z., Ke, R., Pu, Z., Ma, X., & Wang, Y. (2020). Learning traffic as a graph: A gated graph wavelet recurrent neural network for network-scale traffic prediction. *Transportation Research Part C: Emerging Technologies*, *115*, 102620.

Cupertino, T. H., Huertas, J., & Zhao, L. (2013). Data clustering using controlled consensus in complex networks. *Neurocomputing*, *118*, 132–140.

Dai, X., Fu, R., Zhao, E., Zhang, Z., Lin, Y., Wang, F. Y., & Li, L. (2019). DeepTrend 2.0: A light-weighted multi-scale traffic prediction model using detrending. *Transportation Research Part C: Emerging Technologies*, *103*, 142–157.

Defferrard, M., Bresson, X., & Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems*, 3844–3852.

Dharia, A., & Adeli, H. (2003).Neural network model for rapid forecasting of freeway link travel time. *Engineering Applications of Artificial Intelligence*, *16*(7–8), 607–613.

Do, L. N. N., Vu, H. L., Vo, B. Q., Liu, Z., & Phung, D. (2019). An effective spatial-temporal attention based neural network for traffic flow prediction. *Transportation Research Part C: Emerging Technologies*, *108*, 12–28.

Ermagun, A., & Levinson, D. (2019). Spatiotemporal short-term traffic forecasting using the network weight matrix and systematic detrending. *Transportation Research Part C: Emerging Technologies*, *104*, 38–52.

Feng, X., Ling, X., Zheng, H., Chen, Z., & Xu, Y. (2019). Adaptive multi-kernel SVM with spatial–temporal correlation for short-term traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, *20*(6), 2001–2013.

Ganji, A., Shekarrizfard, M., Harpalani, A., Coleman, J., & Hatzopoulou, M. (2020). Methodology for spatio-temporal predictions of traffic counts across an urban road network and generation of an on-road greenhouse gas emission inventory. *Computer-Aided Civil and Infrastructure Engineering*, *35*(10), 1063–1084.

Ghosh-dastidar, S., Adeli, H., & Asce, H. M. (2006). Neural network-wavelet microsimulation model for delay and queue length estimation at freeway work zones. *Journal of Transportation Engineering*, *132*(4), 331–341.

Gu, Y., Lu, W., Qin, L., Li, M., & Shao, Z. (2019). Short-term prediction of lane-level traffic speeds: A fusion deep learning model. *Transportation Research Part C: Emerging Technologies*, *106*, 1–16.

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

Guo, S., Lin, Y., Feng, N., Song, C., & Wan, H. (2019). Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 922–929.

Hamner, B. (2010). Predicting travel times with context-dependent random forests by modeling local and aggregate traffic flow. *2010 IEEE International Conference on Data Mining Workshops* (pp. 1357–1359). IEEE.

Hashemi, H., & Abdelghany, K. (2018). End-to-end deep learning methodology for real-time traffic network management. *Computer-Aided Civil and Infrastructure Engineering*, *33*(10), 849–863.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). IEEE.

Huang, W., Song, G., Hong, H., & Xie, K. (2014). Deep architecture for traffic flow prediction: Deep belief networks with multitask learning. *IEEE Transactions on Intelligent Transportation Systems*, *15*(5), 2191–2201.

Jiang, X., & Adeli, H. (2004). Wavelet packet-autocorrelation function method for traffic flow pattern analysis. *Computer-Aided Civil and Infrastructure Engineering*, *19*(5), 324–337.

Jiang, X., Adeli, H., & Asce, H. M. (2005). Dynamic wavelet neural network model for traffic flow forecasting. *Journal of Transportation Engineering*, *131*(10), 771–779.

Kingma, D. P., & Ba, J. L. (2014). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015–Conference Track Proceedings*. https://arxiv.org/pdf/1412.6980.pdf

Kipf, T. N, & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*. arXiv:1609.02907v4

Kumar, S. V. (2017). Traffic flow prediction using Kalman filtering technique. *Procedia Engineering*, *187*, 582–587.

Lee, S., & Fambro, D. B. (1999). Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting. *Transportation Research Record*, 1678, 179–188.

Li, L., Chen, X., Li, Z., & Zhang, L. (2013). Freeway travel-time estimation based on temporal–spatial queueing model. *IEEE Transactions on Intelligent Transportation Systems*, *14*(3), 1536–1541.

Li, L., Qin, L., Qu, X., Zhang, J., Wang, Y., & Ran, B. (2019). Day-ahead traffic flow forecasting based on a deep belief network optimized by the multi-objective particle swarm algorithm. *Knowledge-Based Systems*, *172*, 1–14.

Li, W., Wang, J., Fan, R., Zhang, Y., Guo, Q., Siddique, C., & Ban, X. J. (2020).Short-term traffic state prediction from latent structures: accuracy vs. efficiency. *Transportation Research Part C: Emerging Technologies*, *111*, 72–90.

Li, Y., Yu, R., Shahabi, C., & Liu, Y. (2017). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *International Conference on Learning Representations*. https://arxiv.org/abs/1707.01926v3

Liu, L., Qiu, Z., Li, G., Wang, Q., Ouyang, W., & Lin, L. (2019). Contextualized spatial–temporal network for taxi origin-destination demand prediction. *IEEE Transactions on Intelligent Transportation Systems*, *20*(10), 3875–3887.

Liu, Y., Liu, Z., Vu, H. L., & Lyu, C. (2020). A spatio-temporal ensemble method for large-scale traffic state prediction. *Computer-Aided Civil and Infrastructure Engineering*, *35*(1), 26–44.

Lv, Y., Duan, Y., Kang, W., Li, Z., & Wang, F.-Y. (2014). Traffic flow prediction with big data: a deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, *16*(2), 865–873.

Ma, X., Tao, Z., Wang, Y., Yu, H., & Wang, Y. (2015). Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C: Emerging Technologies*, *54*, 187–197.

Min, X., Hu, J., Chen, Q., Zhang, T., & Zhang, Y. (2009). Short-term traffic flow forecasting of urban network based on dynamic STARIMA model. *2009 12th International IEEE Conference on Intelligent Transportation Systems* (pp. 1–6). IEEE.

Okutani, I., & Stephanedes, Y. J. (1984). Dynamic prediction of traffic volume through Kalman filtering theory. *Transportation Research Part B*, *18*(1), 1–11.

Pan, Z., Liang, Y., Wang, W., Yu, Y., Zheng, Y., & Zhang, J. (2019). Urban traffic prediction from spatio-temporal data using deep meta learning. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *1*, 1720–1730.

Park, C., Lee, C., Bahng, H., Tae, Y., Jin, S., Kim, K., Ko, S., & Choo, J. (2020). ST-GRAT: A novel spatio-temporal graph attention networks for accurately forecasting dynamically changing road speed. *International Conference on Information and Knowledge Management, Proceedings* (pp. 1215–1224). Association for Computing Machinery

Pereira, D. R., Piteri, M. A., Souza, A. N., Papa, J. P., & Adeli, H. (2020). FEMa: a finite element machine for fast learning. *Neural Computing and Applications*, *32*(10), 6393–6404.

Prigogine, I., Herman, R., & Chaiken, J. (1972). Kinetic theory of vehicular traffic. *Physics Today*, *25*(2), 56–57.

Rafiei, M. H., & Adeli, H. (2017). A new neural dynamic classification algorithm. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(12), 3074–3083.

Rao, W., Wu, Y. J., Xia, J., Ou, J., & Kluger, R. (2018). Origin-destination pattern estimation based on trajectory reconstruction using automatic license plate recognition data. *Transportation Research Part C: Emerging Technologies*, *95*, 29–46.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference*, *57*, 61.

Sibson, R. (1973). SLINK: An optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, *16*(1), 30–34.

Tang, J., Chen, X., Hu, Z., Zong, F., Han, C., & Li, L. (2019a). Traffic flow prediction based on combination of support vector machine and data denoising schemes. *Physica A: Statistical Mechanics and Its Applications*, *534*, 120642.

Tang, J., Li, L., Hu, Z., & Liu, F. (2019b). Short-term traffic flow prediction considering spatio-temporal correlation: A hybrid model combing type-2 fuzzy C-means and artificial neural network. *IEEE Access*, *7*, 101009–101018.

Tang, J., Liu, F., Zou, Y., Zhang, W., & Wang, Y. (2017). An improved fuzzy neural network for traffic speed prediction considering periodic characteristic. *IEEE Transactions on Intelligent Transportation Systems*, *18*(9), 2340–2350.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems* (pp. 5998–6008). NY Curran Associates, Inc.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. *International Conference on Learning Representations*. https://arxiv.org/abs/1710.10903

Wang, H., Liu, L., Dong, S., Qian, Z., & Wei, H. (2016). A novel work zone short-term vehicle-type specific traffic speed prediction model through the hybrid EMD–ARIMA framework. *Transportmetrica B*, *4*(3), 159–186.

Wang, M., Zheng, D., Ye, Z., Gan, Q., Li, M., Song, X., Zhou, J., Ma, C., Yu, L., Xiao, T., He, T., Karypis, G., Li, J., & Zhang, Z. (2019). Deep graph library: a graph-centric, highly-performant package for graph neural networks, 1–7. http://arxiv.org/abs/1909.01315

Wang, X., He, X., Cao, Y., Liu, M., & Chua, T.-S. (2019). KGAT: Knowledge graph attention network for recommendation. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 950–958). Association for Computing Machinery

Williams, B. M., & Hoel, L. A. (2003). Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. *Journal of Transportation Engineering*, *129*(6), 664–672.

Wu, Y. J., Chen, F., Lu, C. T., & Yang, S. (2016). Urban traffic flow prediction using a spatio-temporal random effects model. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, *20*(3), 282–293.

Yan, Y., Zhang, S., Tang, J., & Wang, X. (2017). Understanding characteristics in multivariate traffic flow time series from complex network structure. *Physica A: Statistical Mechanics and Its Applications*, *477*, 149–160.

Yao, B., Chen, C., Cao, Q., Jin, L., Zhang, M., Zhu, H., & Yu, B. (2017). Short-term traffic speed prediction for an urban corridor. *Computer-Aided Civil and Infrastructure Engineering*, *32*(1), 154–169.

Zhang, Y., Cheng, T., & Ren, Y. (2019). A graph deep learning method for short-term traffic forecasting on large road networks. *Computer-Aided Civil and Infrastructure Engineering*, *34*(10), 877–896.

Zhang, Y., Haghani, A., & Zeng, X. (2015). Component GARCH models to account for seasonal patterns and uncertainties in travel-time prediction. *IEEE Transactions on Intelligent Transportation Systems*, *16*(2), 719–729.

Zhang, Z., Li, M., Lin, X., Wang, Y., & He, F. (2019).Multistep speed prediction on traffic networks: A deep learning approach considering spatio-temporal dependencies. *Transportation Research Part C: Emerging Technologies*, *105*, 297–322.

Zhang, L., Liu, Q., Yang, W., Wei, N., & Dong, D. (2013). An improved K-nearest neighbor model for short-term traffic flow prediction. *Procedia–Social and Behavioral Sciences*, *96*, 653–662.

Zhang, J., Shi, X., Xie, J., Ma, H., King, I., & Yeung, D. (2018). GaAN: Gated attention networks for learning on large and spatiotemporal graphs. *UAI 2018: The Conference on Uncertainty in Artificial Intelligence (UAI)* (pp. 339–349). NY Curran Associates, Inc.

Zhang, Y., Smirnova, M. N., Bogdanova, A. I., Zhu, Z., & Smirnov, N. N. (2018). Travel time estimation by urgent-gentle class traffic flow model. *Transportation Research Part B: Methodological*, *113*, 121–142.

Zhang, C., Yu, J. J. Q., & Liu, Y. (2019). Spatial-temporal graph attention networks: A deep learning approach for traffic forecasting. *IEEE Access*, *7*, 166246–166256.

Zhang, J., Zheng, Y., Sun, J., & Qi, D. (2020). Flow prediction in spatio-temporal networks based on multitask deep learning. *IEEE Transactions on Knowledge and Data Engineering*, *32*(3), 468–478.

Zhang, J., Zheng, Y., Qi, D., Li, R., Yi, X., & Li, T. (2018). Predicting citywide crowd flows using deep spatio-temporal residual networks. *Artificial Intelligence*, *259*, 147–166.

Zhang, S., Zhou, L., Chen, X., Zhang, L., Li, L., & Li, M. (2020). Network-wide traffic speed forecasting: 3D convolutional neural network with ensemble empirical mode decomposition. *Computer-Aided Civil and Infrastructure Engineering*, *35*(10), 1132–1147.

Zhao, L., Song, Y., Zhang, C., Liu, Y., Wang, P., Lin, T., Deng, M., & Li, H. (2019). T-GCN: A temporal graph convolutional network for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems*, *21*(9), 3848–3858.

Zheng, C., Fan, X., Wen, C., Chen, L., Wang, C., & Li, J. (2020). DeepSTD: Mining spatio-temporal disturbances of multiple context factors for citywide traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, *21*(9), 3744–3755.

Zou, Y., Hua, X., Zhang, Y., & Wang, Y. (2015). Hybrid short-term freeway speed prediction methods based on periodic analysis. *Canadian Journal of Civil Engineering*, *42*(8), 570–582.

Cheng, X., Zhang, R., Zhou, J., & Xu, W.(2018). DeepTransport: Learning Spatial-Temporal Dependency for Traffic Condition Forecasting. *2018 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Zheng C., Fan X., Wang C., Qi J. (2020). GMAN: A Graph Multi-Attention Network for Traffic Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*, (01), 1234–1241. https://doi.org/10.1609/aaai.v34i01.5477.

Adeli, H., & Karim, A.(2005). *Wavelets in Intelligent Transportation Systems*, John Wiley & Sons, Inc..

How to cite this article: Tang J, Zeng J. Spatiotemporal gated graph attention network for urban traffic flow prediction based on license plate recognition data. *Comput Aided Civ Inf*. 2021;1–21. https://doi.org/10.1111/mice.12688