2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)
October 8-12, 2022, Macau, China

Asymmetric Long-Term Graph Multi-Attention Network for Traffic Speed Prediction

Jiyoung Hwang[1], Byeongjoon Noh[2], Zhixiong Jin[3], and Hwasoo Yeo[4]

2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC) | 978-1-6654-6880-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ITSC55140.2022.9922130

Abstract— **Traffic speed prediction is essential for efficient traffic operation and management by distributing demand concentration in time and space. To make an accurate prediction, it is required to consider spatio-temporal characteristics of the traffic evolution. Recently, deep learning-based approaches, especially Graph Neural Network (GNN) has been widely adopted to reflect the stated characteristics. However, existing GNN models mainly used for short-term prediction, whereas long-term traffic prediction is more useful by enabling earlier and efficient decisions of traffic management as well as individual travels. In this study, we propose Asymmetric Long-Term Graph Multi-Attention Network (ALT-GMAN) algorithm, an extension of the GMAN. ALT-GMAN can predict short and long-term traffic speed by considering asymmetric characteristics of forward and backward waves observed in real roadways. ALT-GMAN is tested with six months highway data of PeMS-Bay area, and MAPE for 3-hours and 6-hours prediction is evaluated as 5.53% and 6.05%, respectively. ALT-GMAN outperforms the existing models in short-term speed prediction, and provides a robust performance in long-term prediction problems, too.**

I. INTRODUCTION

Road traffic speed prediction has been one of the attentive issues in traffic operation and management. An accurate traffic prediction becomes significant because it can help relieve traffic congestion problems and suggest beneficial strategies for traffic applications. It is also useful for dispatching and ride-sharing; the efficient operation and assignment for taxi drivers and optimizing the balanced travel time among candidates for vehicle-sharing or routing strategy by using urban mobility data; the distributed Origin-Destination (OD) pairs [1].

The speed prediction can be classified into two approaches: short-term and long-term prediction. We regard predictions as a short-term prediction when the prediction horizon is less than or equal to 1-hour, while the long-term prediction is to forecast the future greater than 1-hour. Most of the existing models belong to short-term prediction, usually having 15 min and 30 min prediction horizons. However, long-term traffic prediction is also an important topic not only for individual travel, but also for traffic management as it can help travelers to determine departure time and path. Future information to a destination for long-distance can help drivers to make earlier and decisions to avoid unnecessary congestion and accident risk. An accurate traffic speed prediction provides traffic management agencies the information to optimize traffic flow for social cost savings[4].

Traffic speed prediction problem has been traditionally approached using time-series models. For example, [5] used the Autoregressive Integrated Moving Average (ARIMA) model. But it could not efficiently consider spatial features of road networks resulting in the difficulty in handling the complex non-linear relationships among spatio-temporal features [3][6]. We expect that Machine Learning (ML) including Deep Learning (DL) techniques could address those challenges. However, the problem is how to incorporate the real traffic situations and phenomena in the models.

We can represent the road networks with spatio-temporal information by the graph structure [6] because it is useful to handle the connectivity of each road segment and topological structure, and can introduce a graph neural network (GNN) which has shown remarkable performances in graph-related tasks. However, existing graph-based prediction methods use data from all components with less relation to the future traffic situation, and these lead to lower efficiency in learning process for speed prediction. For example, Graph Multi-Attention Network [17] calculates the attention values of all components in graph. Therefore, it is important to use data that is relevant to future traffic conditions for efficiency. We can modify and improve the existing GNN-based model to reflect the knowledge from the traffic theory which can interpret different traffic states of free-flow and congestion, and transitional states of shock waves propagation.

Meanwhile, the long-term prediction by applying existing short-term prediction methods and feature compositions shows increasing errors with the prediction horizon. Several researchers tried to predict long-term traffic with prediction horizons for days or weeks, but prediction horizons to the hours more than 1-hour to several hours have not been studied much although they are the most practical one to travelers and traffic managers. Some existing long-term forecasting methods use the repeatability or the expected similarity with the historical data [4][11]. A survey of traffic speed prediction[2] shows that prediction more than 1-hour requires much more extended input time sequence data. The amount of data required for the long-term prediction can be a weakness of the approach and therefore, this paper tries to suggest the long-term prediction methodology whose

[1]Jiyoung Hwang is with the Civil and Environmental Engineering Department, Korea Advanced Institute of Science and Technology, Daejeon, 34141, South Korea. hjy8185@kaist.ac.kr

[2]Byeongjoon Noh is with the Applied Science Research Institute, Korea Advanced Institute of Science and Technology, Daejeon, 34141, South Korea. powernoh@kaist.ac.kr

[3]Zhixiong Jin is with the Civil and Environmental Engineering Department, Korea Advanced Institute of Science and Technology, Daejeon, 34141, South Korea. iziz56@kaist.ac.kr

[4]Hwasoo Yeo is with the Civil and Environmental Engineering Department, Korea Advanced Institute of Science and Technology, Daejeon, 34141, South Korea. hwasoo@kaist.ac.kr

required data is relatively small and still enough to accurately extract long-term traffic trends.

To address these challenges, we propose a new approach using Graph Neural Network (GNN)-based model: Asymmetric Long-Term Graph Multi-Attention Network (ALT-GMAN). The main points of the proposed model lie on the considerations of the asymmetric characteristics of traffic flow. It extracts the highly related data in model construction and has a small data size to capture long-term dependency. Traffic flow can be identified as a free-flow or congestion state in a flow-density plot. To be specific, each state has different waves(forwards and backwards) with different propagation speeds and influential ranges, and these differences in the two states make the asymmetric structure [12]. We consider this asymmetry concept of traffic flow theory to reflect traffic flow dynamics in speed prediction.

II. LITERATURE REVIEW

A. Traffic Speed Prediction Method

Many statistical models have been developed and applied in traffic speed prediction including ARIMA[5]. However, classical statistical models have shown limited performances. With the development of machine learning methods, we can expect that the new approach can address complex traffic prediction problems with spatial and temporal dependencies, and non-linearity. RNN models have been used successfully for time-series data prediction problems and outperformed the conventional statistical models. As an attempt to capture spatial dependencies, the authors in [7] adopted a CNN architecture by building image-like traffic speed data.

Recently, GNN-based approaches have appeared as the graph structure can efficiently represent road networks with spatial connectivity information. Additionally, Graph Convolutional Network (GCN) is also widely used in many researches [9][10][14]. For example, spatio-temporal Graph Convolutional Network (STGCN) suggests the extraction of meaningful features and patterns in the space domain from graph CNN and the extraction of temporal features from gated CNN. The Dynamic graph convolution (DGC) module in [13] learns the upstream-downstream asymmetric tendency from the spatio-temporal graph and dynamic graph convolution. Among GNN-based techniques, the combined model shows remarkable performances in predicting traffic speed [15][16]. In recent years, attention algorithms that have been used in the analysis of sentence contexts, are applied to traffic prediction problems. The attention algorithm can show the influencing parts of the graph on the target node with higher performance. So, this algorithm embedded in GNN is considered as the state-of-the-art of existing traffic speed prediction models.

B. Graph Construction

A typical GNN model requires a well-defined node and edge structure. Adjacency matrices of a graph represent the features and relationship between each node and edge. Graph adjacency matrices are categorized into mainly four types: road-based, distance-based, similarity-based, and dynamic [6]. Existing GNN speed prediction models mostly used road-based and distance-based ones. Some existing ones used distance-based matrices, and applied the normalized distances into matrices. The authors in [9][17] calculated all the distances between nodes, but this idea has disadvantages. Even if the distance between two nodes is close, each node can be irrelevant without considering the directions of traffic flow. the future speed can be influenced by many features such as flow, future traffic demand from upstream sections connected, and existing queueing situation from the downstream sections. Therefore, to construct a well-defined adjacency matrix and graph network, we should consider the road characteristics and direction. Notice that Google Deepmind [18] suggested a super-segment concept to predict the speed of each segment considering the influences from adjacent sections. One research [19] shows different attentions on upstream and downstream according to congestion spreading and congestion dissipating. In specific, in the situation of congestion spreading, the wave is transferred to backward (to upstream) vehicles so that the downstream has much attention weight. The circumstance of congestion dissipating shows that upstream has much attention weight. The above traffic engineering knowledge-based embedding is essential because it deals with actual road condition data. Consequently, appropriate graph structure and embedding process should contain traffic engineering basics and traffic flow theory.

III. METHODOLOGY

A. Data Description

In this study, the proposed model is evaluated with the Caltrans Performance Measurement System (PeMS)-Bay area data [1][2][6]. It contains the highway traffic speed from January 1st, 2017 to June 30th, 2017. The data includes 5 minutes timestep speed data from 321 sensors installed on 7 highways in the Santa Clara region. PeMS-Bay area data is a widely-used baseline dataset to evaluate the performance of GNN models [1][2][6][15]. The sensor distributions are expressed in Fig 1. To consider each roadway's connectivity and movement of traffic flow, we classify and arrange the sensor into each road link unit according to the absolute distance from the datum point. Each sensor responds to each vertex of GNN. We split the data; 70% for training, 10% for validation, and 20% for testing.

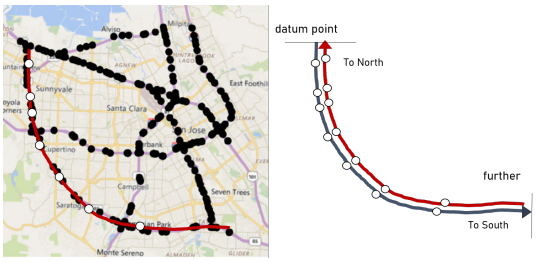

Fig. 1. PeMS-Bay area data description

B. Framework

The entire framework of the proposed ALT-GMAN is described in Fig 4 containing data collection & process, spatial embedding, long-term temporal embedding, and prediction procedures. We defined $G = (V, E, A)$ as a spatial graph network with a static directed connection. V indicates the set of vertices, $|V| = N$ (N indicates the number of vertices). E is the set of edges that shows the connectivity between vertices. $A \in R^{N \times N}$ represents the adjacency matrix of the G and A indicates the connection between two nodes v_i and v_j. If $v_i, v_j \in V$ and $(v_i, v_j) \in E$, A is one, otherwise, it is zero. Traffic condition information is represented as a graph signal $X_t \in R^{N \times C}$ on graph G, where C denotes the number of traffic conditions. In this paper, traffic condition is traffic speed information detected from road-side sensors. A road network consists of N nodes, and each node generates a time series of traffic speed data. We extracted historical P time steps and the observed past time sequence data is represented by row vectors.

$$\chi = (X_{t_1}, X_{t_2}, \ldots, X_{t_P}) \in R^{P \times N \times C} \qquad (1).$$

,Our goal is to forecast the traffic speed of the next Q time steps for all sensor nodes. The formulation is below

$$\hat{Y} = (\hat{X}_{t_{P+1}}, \hat{X}_{t_{P+2}}, \ldots, \hat{X}_{t_{P+Q}}) \in R^{Q \times N \times C} \qquad (2)$$

where,
X_{t_i} = historical road link speed of time step i
$\hat{X}_{t_{P+i}}$ = predicted road link speed of time step i
χ = historical sequence of road link speed data
$\hat{Y}$ = predicted sequence of road link speed data.

C. Graph Multi-Attention Network (GMAN)

The previously developed GMAN [17] algorithm shows remarkable performances in predicting traffic speed among GNN-based techniques [6]. GMAN uses three main structures encoder-decoder, spatio-temporal embedding (STE), and graph multi-attention block. Additionally, there is a transformer attention layer and a fully connected (FC) layer for both encoder and decoder. GMAN structure has residual connections with STE which contain spatial attention and temporal attention with gated fusion. A transformer attention layer converts the encoded time sequence data into the decoded one. Graph structure and temporal information are incorporated into multi-attention. In the spatial embedding process, vertex and connection information are expressed in random walk-based node embedding, and temporal embedding is added to represent dynamic correlations among traffic sensors. Temporal embedding has two types; day-of-week and time-of-day, and these features are encoded. GMAN implemented the short-term prediction; 15-min, 30-min, and 1-hour.

D. Traffic Characteristics

Roadway traffic can be simply characterized by the fundamental diagram that shows the relationship between flow(q) and density(k). The fundamental diagram shows free-flow speed and wave speed from a macroscopic view. Traffic can be classified into two states: (1) free-flow state in which vehicles can move at desired speed without influence by downstream traffic conditions, and (2) congestion state with slow speed and have high influence by the downstream traffic. Fig 2 shows a typical triangular-shape fundamental diagram. Free-flow speed denotes the speed of vehicles in the free-flow state. In the congestion situation, the congestion propagates in the opposite direction with wave speed. By considering the difference between two traffic states, the influential areas of upstream and downstream traffic can be regarded as asymmetric. These asymmetric influential area concepts are introduced in the next section and Fig 3.

E. Spatial Embedding

Spatial embedding (SE) is a necessary step to contain limited information on the underlying road network. As mentioned in the previous section, the well-defined graph construction is the basis of SE. We constructed an adjacency matrix with the classified nodes depending on stream types. We set the adjacency matrix as a directed connection matrix. As vehicles from upstream arrive with free-flow speed(forward), and the congestion wave from downstream propagates with the wave speed(backward), the free-flow state is influenced by upstream sections and the congestion state by downstream ones. The observation range (node unit) of upstream and downstream are denoted as Upstream Influential Area (UIA) and Downstream Influential Area (DIA), respectively. The notion of UIA and DIA implies asymmetric traffic characteristics, and it represents the over-

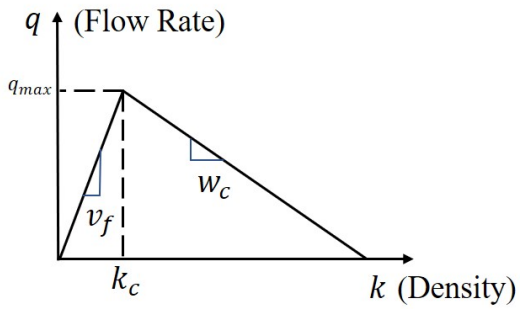

Fig. 2. Fundamental diagram

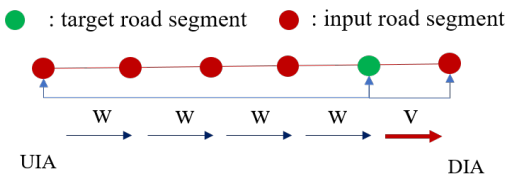

Fig. 3. Asymmetric UIA and DIA

all traffic situation on the roadway like backward propagation of congestion shock waves. The example of asymmetric UIA and DIA is described in Fig 3, and values of UIA and DIA are determined by the reciprocal ratio of the free-flow speed and the wave speed as in equation 3.

$$v_f = \frac{UIA}{DIA} w_c \tag{3}.$$

The applied asymmetric structures on the adjacency matrix with UIA and DIA are shown in ALT-STE part of Fig 5, and it shows the entire structure of ALT-GMAN. The left column of Fig 4 shows SE either. This spatial embedded file contains direction and connection information and asymmetric characteristics between each node.

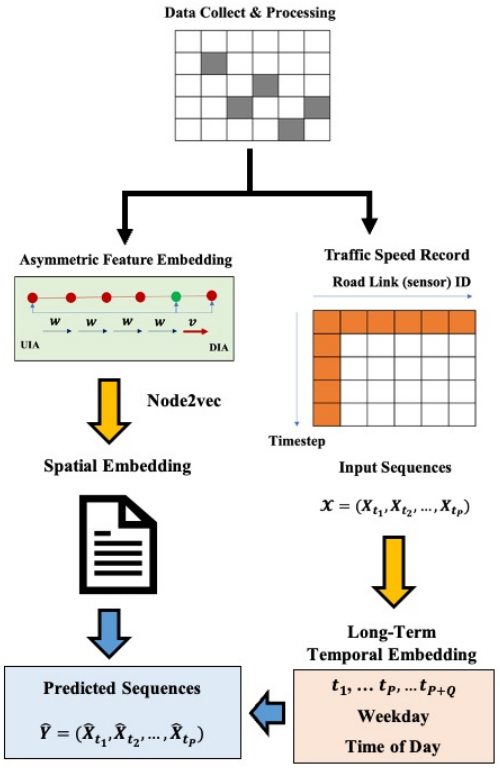

Fig. 4. ALT-GMAN Framework

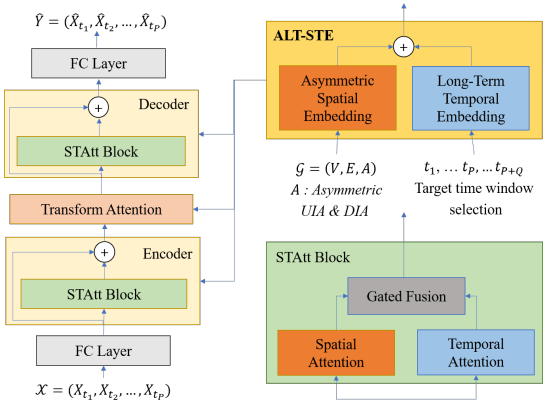

Fig. 5. ALT-GMAN Structure

F. Temporal Embedding

Temporal embedding (TE) processes are based on the idea that a long-term prediction model can have the ability to predict several hours. For this, We extracted 12-time steps from the historical data sequence and set 12-time steps as the prediction horizon. The model is supposed to learn the tendency and relation between the current traffic situation and the future one. We implemented 3-hours, 6-hours of long-term predictions. Additionally, we implemented 15-min, and 30-min short-term predictions by varying TE into short-term. These procedures are also shown in the ALT-STE part of Fig 5 and the right column of Fig 4.

IV. EXPERIMENTS

A. Experiment design

We applied three metrics for evaluation purposes: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). In the GMAN model, there are three hyperparameters: the number of ST-Attention blocks L, the number of attention heads K, and the dimensionality d of each attention head. The channel of each layer D equals the multiplication of K and d. We set the batch size as 32, and we used part of these parameters from the original model. We used $L = 1, K = 8, d = 8 (D = 64)$.

In our experiment, we constructed two experiments, applying only SE, and STE. We also compared ALT-GMAN with the following baseline methods: ARIMA [5], SVR, Feed-forward neural network (FNN), FC-LSTM, STGCN [13], DCRNN [9], Graph WaveNet [10], and GMAN [17]. For models ARIMA, SVR, FNN, and FC-LSTM, we used the

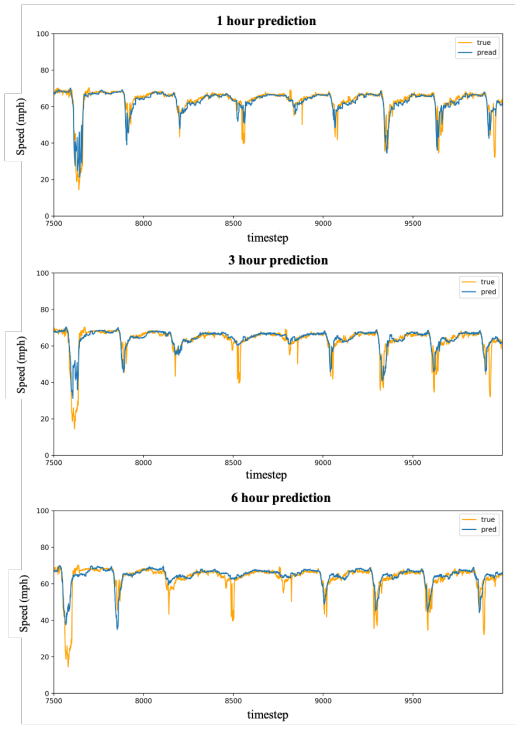

Fig. 6. ALT-GMAN performance - PeMS-Bay

settings of [9]. For models STGCN, DCRNN, and Graph WaveNet [10], we used the default settings of their original proposals.

V. RESULTS

We validated the feasibility and applicability of the proposed model by experiments of short and long-term speed predictions with considering asymmetric traffic flow characteristics.

A. Short-Term Traffic Speed Prediction

In order to measure the prediction performance, we implemented the short-term traffic prediction with Asymmetric GMAN. GMAN embedded with only SE technique is defined as AGMAN and TE is not used in the case of short-term prediction. We compared AGMAN with the above baseline methods. TABLE I shows the comparison result of different methodologies for short-term prediction: 15-min, 30-min, and 1-hour prediction horizon. We observed that DL-based techniques outperform traditional time series-based approaches and ML models. Moreover, the GNN-based approaches show remarkable performance improvement than the CNN and RNN models as GNN models contain the road network information, and they can consider the connection among nodes. As described in TABLE I, the performance of the AGMAN shows the best performance among the baselines. The SE technique, which applies the asymmetric structure and different influential areas in traffic flow would make enhanced performances.

B. Long-Term Traffic Speed Prediction

The TE technique enables the long-term prediction more than 1-hour. We implemented prediction on 1-hour, 3-hours, and 6-hours time horizons (see TABLE II). Error propagation effect decreases much more than the case with predicting long-term traffic indirectly through short-term prediction by applying TE methodology. Moreover, Fig 6 shows that the ALT-GMAN model can capture the timing of speed drop when congestion occurs and the timing of speed recovery. It means that ALT-GMAN has learned long-term traffic trends from past data based on the asymmetric traffic behaviors. Consequently, TE enables long-term prediction and SE would make the enhanced performance.

VI. CONCLUSIONS

In this study, we proposed ALT-GMAN that predicts long-term and short-term traffic speeds. The main contributions of this study lie in (1) embedding of the concept of asymmetric characteristics of traffic flow depending on the state in the learning process, and (2) the proposed temporal embedding technique that requires a small data size to capture a long-term tendency.

It should be noted that this study is motivated by a lack of efficient long-term speed prediction models considering spatio-temporal features of the road network and asymmetric characteristics of the traffic flow. ALT-GMAN outperformed in forecasting both short and long-term speeds, and it can support making efficient decisions in traffic management and individual travel.

ACKNOWLEDGMENT

This work was supported by Korea Institute of Police Technology (KIPoT) grant funded by the Korea government (KNPA) (092021C29S01000, Development of Traffic Congestion Management System for Urban Network).

REFERENCES

[1] H. Yuan and G. Li, "A Survey of Traffic Prediction: from Spatio-Temporal Data to Intelligent Transportation," Data Sci. Eng., no. 0123456789, 2021, doi: 10.1007/s41019-020-00151-z.

[2] D. A. Tedjopurnomo, Z. Bao, B. Zheng, F. Choudhury, and A. K. Qin, "A Survey on Modern Deep Neural Network for Traffic Prediction: Trends, Methods and Challenges," IEEE Trans. Knowl. Data Eng., vol. 14, no. 8, pp. 1–1, 2020, doi: 10.1109/tkde.2020.3001195.

[3] M. G. Karlaftis and E. I. Vlahogianni, "Statistical methods versus neural networks in transportation research: Differences, similarities and some insights," Transp. Res. Part C Emerg. Technol., vol. 19, no. 3, pp. 387–399, 2011, doi: 10.1016/j.trc.2010.10.004.

[4] L. E. Architecture, Z. Wang, X. Su, and Z. Ding, "Long-Term Traffic Prediction Based on," Ieee Trans. Intell. Transp. Syst., pp. 1–11, 2020.

[5] S. Makridakis and M. Hibon, "ARMA models and the Box-Jenkins methodology," J. Forecast., vol. 16, no. 3, pp. 147–163, 1997, doi: 10.1002/(SICI)1099-131X(199705)16:3¡147::AID-FOR652¿3.0.CO;2-X.

[6] W. Jiang and J. Luo, "Graph Neural Network for Traffic Forecasting: A Survey," vol. 14, no. 8, pp. 1–19, 2021, [Online]. Available: http://arxiv.org/abs/2101.11174.

[7] Y. Kim, P. Wang, Y. Zhu, and L. Mihaylova, "A capsule network for traffic speed prediction in complex road networks," arXiv, pp. 6–11, 2018.

[8] Z. Lv, J. Xu, K. Zheng, H. Yin, P. Zhao, and X. Zhou, "LC-RNN: A deep learning model for traffic speed prediction," IJCAI Int. Jt. Conf. Artif. Intell., vol. 2018-July, pp. 3470–3476, 2018, doi: 10.24963/ijcai.2018/482.

[9] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," 6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc., pp. 1–16, 2018.

[10] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, "Graph wavenet for deep spatial-temporal graph modeling," IJCAI Int. Jt. Conf. Artif. Intell., vol. 2019-Augus, pp. 1907–1913, 2019, doi: 10.24963/ijcai.2019/264.

[11] Z. Hou and X. Li, "Repeatability and Similarity of Freeway Traffic Flow and Long-Term Prediction under Big Data," IEEE Trans. Intell. Transp. Syst., vol. 17, no. 6, pp. 1786–1796, 2016, doi: 10.1109/TITS.2015.2511156.

[12] H. Yeo, Asymmetric Microscopic Driving Behavior Theory, no. May. 2008.

[13] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," IJCAI Int. Jt. Conf. Artif. Intell., vol. 2018-July, pp. 3634–3640, 2018, doi: 10.24963/ijcai.2018/505.

[14] G. Li, V. L. Knoop, and H. van Lint, "Multistep traffic forecasting by dynamic graph convolution: Interpretations of real-time spatial correlations," Transp. Res. Part C Emerg. Technol., vol. 128, no. May, p. 103185, 2021, doi: 10.1016/j.trc.2021.103185.

[15] L. Zhao et al., "T-GCN: A Temporal Graph Convolutional Network for Traffic Prediction," IEEE Trans. Intell. Transp. Syst., vol. 21, no. 9, pp. 3848–3858, 2020, doi: 10.1109/TITS.2019.2935152.

[16] Z. Cui, K. Henrickson, R. Ke, and Y. Wang, "Traffic Graph Convolutional Recurrent Neural Network: A Deep Learning Framework for Network-Scale Traffic Learning and Forecasting," IEEE Trans. Intell. Transp. Syst., vol. 21, no. 11, pp. 4883–4894, 2020, doi: 10.1109/TITS.2019.2950416.

[17] C. Zheng, X. Fan, C. Wang, and J. Qi, "GMAN: A graph multi-attention network for traffic prediction," AAAI 2020 - 34th AAAI Conf. Artif. Intell., pp. 1234–1241, 2020, doi: 10.1609/aaai.v34i01.5477.

[18] "Google Deepmind." https://deepmind.com/blog/article/traffic-prediction-with-advanced-graph-neural-networks (accessed Jul. 27, 2021).

TABLE I

COMPARISON OF BASELINE METHODOLOGIES

Data	Model	15 min			30 min			1-hour		
		MAE(mph)	RMSE(mph)	MAPE (%)	MAE(mph)	RMSE(mph)	MAPE (%)	MAE(mph)	RMSE(mph)	MAPE (%)
PeMS	ARIMA	1.62	3.30	3.50	2.33	4.76	5.40	3.38	6.50	8.30
	SVR	1.85	3.59	3.80	2.48	5.18	5.50	3.28	7.08	8.00
	FNN	2.20	4.42	5.19	2.30	4.63	5.43	2.46	4.98	5.89
	FC-LSTM	2.05	4.19	4.80	2.20	4.55	5.20	2.37	4.96	5.70
	STGCN	1.36	2.96	2.90	1.81	4.27	4.17	2.49	5.69	5.79
	DCRNN	1.38	2.95	2.90	1.74	3.97	3.90	2.07	4.74	4.90
	Graph WaveNet	1.30	2.74	2.73	1.63	3.70	3.67	1.95	4.52	4.63
	GMAN	1.34	2.82	2.81	1.42	3.72	3.63	1.86	4.32	4.31
	AGMAN	**1.15**	**2.16**	**2.30**	**1.43**	**2.87**	**3.07**	**1.72**	**3.63**	**3.95**

TABLE II

PERFORMANCE OF ALT-GMAN

Data	Model	Time Horizon (min)	MAE (mph)	RMSE (mph)	MAPE (%)
PeMS	**ALT-GMAN**	15	1.15	2.16	2.30
		30	1.43	2.87	3.07
		60	1.72	3.63	3.95
		180	2.40	5.05	5.53
		360	2.57	5.31	6.05

[19] K. Tang et al., "Short-term Travel Speed Prediction for Urban Expressways: Graph Attention Network Model," in IEEE Transactions on Intelligent Transportation Systems, 2020, pp. 1–12, doi: 10.1109/tits.2020.3027628.