<http://www.diva-portal.org>

Preprint

This is the submitted version of a paper presented at *The 22nd IEEE Intelligent Transportation System Conference*.

Citation for the original published paper:

Zhang, T., Jin, J., Yang, H., Ma, X. (2019)

Link speed prediction for signalized urban traffic network using a hybrid deep learning approach

In:

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-256047>

Short-term prediction of link-based traffic speed for signalized road networks using a hybrid deep learning approach

Tong Zhang, Junchen Jin*, Hui Yang, Haifeng Guo and Xiaoliang Ma

Abstract—Predicting traffic speed is of importance in transportation management. Signalized road networks manifest highly dynamic speed patterns that are challenging to model and predict. We propose an integrated deep-learning based approach for short-term speed prediction, aiming at capturing heterogeneous spatio-temporal correlations between road intersections. After transforming original road networks and intersections into graphs, this approach leverages a layered graph convolution network structure to model traffic speed variations at both intersection and road network levels. The two levels are combined through a fully connected neural layer. Neural spatio-temporal attention mechanisms are applied to modulate the most relevant periodical traffic information during signal cycles. The proposed approach was evaluated using real-world speed data collected in Hangzhou City, China. Experiments demonstrate that the proposed approach can offer a scalable and effective solution for predicting short-term speed for signalized road networks.

I. INTRODUCTION

Predicting short-term traffic speed is of importance in transportation management, such as travel time prediction and traffic signal control. Traditionally, short-term traffic prediction relies on mathematical and statistical modeling approaches, such as autoregressive moving average model [1], Kalman filter [2], [3], and cellular automaton [4]. These analytical approaches are primarily developed based on traffic flow theories, which establish a unified theoretical framework to describe complex spatio-temporal driving behaviors. With the abundance of traffic monitoring data collected from intelligent transportation systems, data-driven machine learning approaches are gaining momentum over the past few years. Data driven approaches are deemed as a prominent solution for traffic modeling due to their prediction performance and adaptability [5], [6]. Among these efforts, deep learning based approaches attract wide attention among both researchers and professionals [7]. Various deep neural networks have been applied to forecast short-term traffic flow, speed, and travel time, such as stacked autoencoders

[8], deep belief network [9], convolutional neural network [10], and long short-term memory neural network [11].

To capture traffic correlations from both spatial and temporal dimensions, researchers have devised advanced network structures that integrate multiple types of deep neural networks into a comprehensive architecture. In particular, the combination of convolutional networks and recurrent networks has been explored extensively [12],[13]. Nevertheless, network-wide prediction usually requires learning complex correlations between the current road link and upstream/downstream links. Meanwhile, local features of each road link and intersection should be accounted for. These considerations pose challenges for the generalization abilities of prediction network models. Moreover, network-wide prediction needs to collect sufficient training data, which are costly and error-prone in realistic applications.

In this study, we address these issues by proposing a two-level integrated deep neural network structure that takes into account both network-wide and local traffic features. In order to model spatio-temporal correlations, we adopt attention mechanisms [14] to automatically identify relevant input data to produce robust prediction outputs. Since road networks can be readily modeled as graphs, we use graph convolutional networks [15],[16] to perform the short-term prediction task due to their advantages in learning spatio-temporal dependencies over arbitrary graphs. The contributions of this study can be summarized as follows:

- We develop an integrated two-level traffic speed prediction method based on graph convolutional neural networks by accounting for both intersection and network traffic dynamics.
- We apply spatio-temporal attention mechanisms to capture spatio-temporal correlations that are crucial for traffic speed modeling within deep neural networks.
- We demonstrate the proposed prediction approach with real-world traffic data collected in Hangzhou City, China.

II. METHODOLOGY

A. Overview

The studied short-term speed prediction problem can be formally formulated as follows. Given historically observed speed data for N individual road links, we aim to predict future speed for these road links within a short period of time (usually less than one hour). In signalized road network, speed information is collected for each signal cycle. We can denote the ending time of a signal cycle as t . The road

Tong Zhang Ph.D. and Hui Yang are with State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (phone: 86-27-68779959; fax: 86-27-68778969; email: zhangt@whu.edu.cn, yangh2017@whu.edu.cn).

Junchen Jin Ph.D. is with Smart Transportation Research Institute, Enjoyor Co. Ltd, Hangzhou, 310030, China and Engineering Research Center of Intelligent Transport of Zhejiang Province, Hangzhou 310030, China.

Haifeng Guo Ph.D. is with the College of Information Engineering, Zhejiang University of Technology, Hangzhou 310013, China. He is also affiliated with Enjoyor Co., Ltd, Hangzhou 310030, China.

Xiaoliang Ma Ph.D. is with Department of Civil and Architectural Engineering, KTH Royal Institute of Technology, Sweden.

*Corresponding author; Email: junchen@kth.se

network can be represented as a directed graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ is a set of road intersections, $\mathbf{E}$ is a set of road links, $|\mathbf{E}| = N$. Each road intersection consists of multiple incoming road links, which in turn comprise multiple lanes (outgoing links are not modeled).

A road intersection can be also represented as a directed graph $\mathbf{G}' = (\mathbf{V}', \mathbf{E}')$, where $\mathbf{V}'$ is a set of road lanes, $\mathbf{E}'$ denotes directed connections from lanes to their downstream lanes. Suppose $s_i(t)$ be a real valued speed observation for road link i and $\mathbf{S}(t) = [s_0(t), \dots, s_i(t), \dots, s_K(t)]$ be a speed measurement vector for the studied road network consisting of K intersections at time t . Given known traffic signal timing plans $\mathbf{TS}$ (including cycle length and ratio of effective green time), the short-term prediction task is to establish a nonlinear function $\mathbf{F}^*$ that maps m past speed observation vectors from the road network $\mathbf{S}(t-m+1), \dots, \mathbf{S}(t-1), \mathbf{S}(t)$ to a predicted speed vector $\mathbf{S}(t+\theta)$ at θ cycles after t (i.e., $t+\theta$) based on $\mathbf{G}$ and $\mathbf{G}'$. The prediction problem is then expressed as

$$\mathbf{F}^* : [\mathbf{S}(t-m+1), \dots, \mathbf{S}(t-1), \mathbf{S}(t); \mathbf{G}, \mathbf{G}', \mathbf{TS}] \rightarrow [\mathbf{S}(t+\theta)] \in \mathcal{R}^N \quad (1)$$

Obviously, there exist complex spatial and temporal correlations between a road link and its neighboring links. For a specific road link, upstream and downstream traffic would exert strong influence on it at the current time t and a few cycles ahead. The temporal effect on the same road link could also be significant. Meanwhile, traffic interactions and dependencies between lanes should be accounted for, despite that they are barely explored in the literature. In order to minimize the error between predicted speed values and actual observations, we ought to consider these spatio-temporal correlations in the proposed prediction model.

The data used in prediction modeling come from three sources: link-level speed records that are derived from floating car data, road network topology and geometry data, and signal timing data for each traffic signal at road intersections.

We construct the prediction model at two levels: network-wide modeling based on graph $\mathbf{G}$ in which road intersections are denoted as nodes and road links are represented as edges; intersection-level modeling that focuses on individual intersections using detailed topological and geometric road network data, as well as signal timing data. The idea behind this scheme is that both network-wide and We construct the prediction model at two levels: network-wide modeling based on graph $\mathbf{G}$ in which road intersections are denoted as nodes and road links are represented as edges; intersection-level modeling that focuses on individual intersections using detailed topological and geometric road network data, as well as signal timing data. The idea behind this scheme is that both network-wide and intersection-level scales are equally critical for accurate prediction, thereby necessitating the integration of traffic dynamics at these two scales.

The workflow of the link-based speed prediction task is presented in Fig. 1. The prediction model consists of two-level models: micro modeling for intersections and network-wide macro modeling. The two models all take the forms of

spatio-temporal graph convolution networks. For each model, we implement spatio-temporal attention mechanisms to filter out irrelevant input information, which is followed by graph convolution operations aiming at extracting traffic features at spatial dimension. Finally the two models are combined via a fully connected layer, which can produce the final prediction results for each incoming road link in the road network.

B. Modeling network-wide traffic

Since we model the entire network as a graph $\mathbf{G}$, detailed geometric information for road links and intersections is disregarded. The speed is aggregated at the interval of 10 minutes, which is much larger than signal cycles. Through the aggregation of both spatial and temporal traffic information at network level, we strive to describe the overall traffic patterns for the network, which can complement the training of spatio-temporal correlations at intersection level.

1) *Spatio-temporal attention*: The idea of attention mechanism is to assign weights for input data so as to describe their contributions to upper-level representations in a more accurate manner [17]. For each neural network layer, suppose we have an input data block $\mathbf{X} \in \mathcal{R}^{K \times T \times F}$, where K is the spatial dimension (number of road intersections), T is the length of temporal dimension, F is the number of input feature channels. We can use a spatial attention model to compute a weighted sum of input data block [18]. A spatial attentional modulation matrix can be written as

$$\mathbf{A}_s = \sigma[(\mathbf{X}\mathbf{W}_t)\mathbf{W}_{t,f}(\mathbf{W}_f\mathbf{X})^T + \mathbf{b}]. \quad (2)$$

where σ is sigmoid activation function, $\mathbf{W}_t \in \mathcal{R}^T$, $\mathbf{W}_{t,f} \in \mathcal{R}^{T \times F}$, $\mathbf{W}_f \in \mathcal{R}^F$, $\mathbf{b} \in \mathcal{R}^{K \times K}$ are parameters to be learned. We omit the index of layer for simplicity. Similarly, a temporal attentional modulation matrix is defined as

$$\mathbf{A}_t = \sigma[(\mathbf{X}^T\mathbf{W}'_t)\mathbf{W}'_{t,f}(\mathbf{W}'_f\mathbf{X})^T + \mathbf{b}']. \quad (3)$$

where $\mathbf{W}'_k \in \mathcal{R}^K$, $\mathbf{W}'_{k,f} \in \mathcal{R}^{K \times F}$, $\mathbf{W}'_f \in \mathcal{R}^F$, $\mathbf{b}'_f \in \mathcal{R}^{K \times K}$ are trainable parameters. Both $\mathbf{A}_s$ and $\mathbf{A}_t$ are further normalized by the softmax function. Then we can apply these attentional modulation matrices to the input data block $\mathbf{X}$ to generate new input data which capture more cohesive spatio-temporal correlations than the original input data. Note $\mathbf{A}_s$ and $\mathbf{A}_t$ are $K \times K$ and $T \times T$ matrices respectively.

2) *Spatial Graph convolution*: Upon the availability of modulated input data based on the above attention mechanisms, we can proceed to apply spatial and temporal graph convolution operations to further capture spatial and temporal correlations embedded in the network-wide traffic dataset. We adopt a commonly used graph convolution operator [19].

$$\mathbf{X}_{l+1} = \sigma(\mathbf{U}\mathbf{g}(\mathbf{A})\mathbf{U}^T\mathbf{X}_l) \quad (4)$$

where $\mathbf{X}_{l+1}$ and $\mathbf{X}_l \in \mathcal{R}^{K \times T \times F}$ are the input data block in layer l and $l+1$ respectively. $\mathbf{U}$ and $\mathbf{A}$ are orthogonal and diagonal matrices respectively. The Laplacian matrix $\mathbf{L}$ of $\mathbf{G}$ can be decomposed as $\mathbf{U}\mathbf{A}\mathbf{U}^T$. $\mathbf{g}_\theta(\cdot)$ is a polynomial kernel function. σ is the Rectified Linear Unit (ReLU) activation function. Therefore, the above operation is essentially conducting convolutions on the orthogonal decomposition

Fig. 1: Workflow of the proposed prediction approach.

of the Laplacian matrix L of G . We first conduct graph convolution operations for the spatial dimension to filter speed information in neighboring intersections, and then do convolution operations on the closest temporal data to capture temporal correlations.

Fig. 2 depicts the architecture of the macro model, which consists of a spatio-temporal modulation module, a spatial graph convolution layer, and recurrent neural network layers. After spatial graph and temporal convolution operations, the model uses node indices to fetch feature information for each node (i.e., intersection) before feeding this information into recurrent neural units. Input data are speed time series with 10-minute interval for all road intersections. The macro model outputs hidden state information for each intersection.

3) *Modeling intersection traffic*: For each road intersection, we can use detailed topological and geometric road data to build a micro training model. These data contains the number of lanes at each constituent road link, lane geometry, lane connectivity information, and turn restrictions. Signal timing information is also leveraged for this micro modelling. We construct graphs for each road intersection by denoting each lane as a graph node and interconnections between these lanes as graph edges. Fig. 3 shows the micro-modelling architecture. The micro model also implements spatio-temporal attention mechanisms and graph convolution operations. The micro model comprises a spatio-temporal attention modulation module, a spatial graph convolution layer, and a temporal graph convolution layer. In the micro model, the spatio-temporal attention and graph convolution operations are defined similarly to those of the macro model. Input data are cycle-based speed time series for a specific road intersection. The micro model outputs high-level abstract feature information for the intersection.

4) *Model integration*: The difference between the two level models is that the network-level model incorporates recurrent network layers after the spatial graph convolution layer, aiming at producing hidden states for each node. At the intersection level, we use a temporal graph convolution layer after the spatial graph convolution layer for the micro model. The hidden state information from the macro model is then fused with temporal convolution operation results from the micro model for the same road intersection. Since

in the macro model, we fetch results for each intersection after graph convolution operations, the integrated feature is produced on the basis of individual intersections. The fusion is done through a linear weighting scheme for the outputs of the two models. A fully-connected layer is then applied to these fused features to generate the final prediction results for each incoming road link that connects to the designated intersection.

III. IMPLEMENTATION

The detailed implementation is summarized as follows.

A. Macro model

To prepare for the macro model, speed data were aggregated into ten minutes interval. The input features for time interval t (i.e., ten minutes) include average speed, maximum speed, minimum speed, whether t falls into morning peak hours, and whether t falls into afternoon peak hours. The first three numeric features are computed based on speed data of the constituent incoming road links of the intersection. The past two-hour data were used to produce input features.

B. Micro model

For the micro model, we directly used data at the original temporal resolutions. The input features for time interval t' (signal cycle length) include average speed, number of lanes, and theratioof effectivegreen time to the cycle length. A graph is constructed for each road intersection to describe topological connections between lanes on different road links. Each intersection is composed of at least 3 links that have incoming lanes. In this study, we only predict for incoming links and ignore outgoing links. Data of the past 30 signal cycles were collected to extract input features.

C. Training settings

We set the sizes of temporal kernel size as 1×3 with a stride of (1×1) . For both macro and micro models, the graph convolution layer has 64 output channels. The number of the fully connected layer is set as the number of incoming links of the intersection to be predicted. The average difference between the predicted value and the actual observed value is used as the loss function. The training was performing by back-propagation, using a batch size of 64 and a learning rate

Fig. 2: Structure of network-level macro model.

Fig. 3: Structure of intersection-level micro model.

Fig. 4: Study region and road network. Intersections being selected for test are marked by star symbols.

of 0.0001. The fusion weight is determined during training. The dimension of prediction output is set to the number of links that connect at the examined intersection. Training terminates after 50 epochs.

D. Data

The proposed prediction approach was evaluated using real-world traffic data collected in Hangzhou City, China.

One month of link-based speed data (Dec. 1 to 31, 2018) were obtained from Autonavi Software Co., Ltd, a leading navigation service company in China. The split ratio of training, validation, and test data is set as 8 : 1 : 1. In addition to the speed data, we have access to traffic signal data and road network data. The signal timing dataset contains cycle length and phrase transition times for each traffic signal. The road network dataset provides detailed topology, geometry and attribute information of lanes. Lane-specific connectivity and turn restriction information are also available in the dataset. The study region consists of 154 road links and 18 intersections (Fig. 4). The road network graph and intersection graphs are represented as adjacent matrices in which each matrix element indicates whether two nodes are connected (one indicates connected and zero otherwise).

E. Experimental results

The proposed prediction approach was compared with two baseline models:

- The graph convolution network model (GCN) without the spatio-temporal attention mechanisms. The goal of this test is to evaluate the contribution of attention modeling to the overall prediction performance.
- A recurrent neural network model (RNN) that only characterizes temporal correlations of traffic data. The

TABLE I: Prediction Performance Comparison for Intersection #767 with Different Prediction Horizons (3, 5, and 10 Cycles Ahead). The Best Results are Shown in Bold.

Prediction Horizon		MAE			RMSE			MAPE (%)		
		3	5	10	3	5	10	3	5	10
Peak hours	RNN	3.70	3.75	3.93	4.62	4.74	5.09	12.18	12.78	14.29
	GCN	3.56	3.55	3.67	4.45	4.49	4.75	11.75	12.16	13.37
	Ours	3.41	3.44	3.63	4.31	4.40	4.73	11.08	11.65	13.19
Non-peak hours	RNN	5.14	5.15	5.27	6.69	6.71	6.82	14.72	14.73	15.02
	GCN	4.83	4.85	4.84	6.35	6.35	6.34	13.75	13.76	13.64
	Ours	4.75	4.79	4.83	6.24	6.26	6.29	13.58	13.66	13.77

TABLE II: Prediction Performance Comparison for Intersection #768 with Different Prediction Horizons (3, 5, and 10 Cycles Ahead). The Best Results are Shown in Bold.

Prediction Horizon		MAE			RMSE			MAPE (%)		
		3	5	10	3	5	10	3	5	10
Peak hours	RNN	9.94	9.19	8.25	12.45	11.48	10.21	31.82	30.42	28.30
	GCN	5.92	5.89	5.91	7.08	7.02	6.97	21.17	20.91	20.97
	Ours	5.63	5.70	5.79	6.71	6.74	6.77	19.87	20.09	20.30
Non-peak hours	RNN	8.80	8.28	7.54	11.34	10.67	9.7	24.86	23.42	21.41
	GCN	5.42	5.41	5.39	6.91	6.93	6.95	16.17	16.10	16.06
	Ours	5.36	5.39	5.40	6.89	6.92	6.95	15.68	15.78	15.78

comparison would reveal the benefits of spatio-temporal attention mechanism and spatial graph convolution operations.

The two baselines models were trained on the historical data of each road intersection using a batch size of 64 and a learning rate of 0.0001. Both models were trained for 50 epochs.

All experiments were conducted on a desktop computer with an Intel Core i7-8700K processor and 32GB memory, using the Windows 7 operating system. The validation and test data were further divided by peak and non-peak hours. These two time intervals exhibit distinct spatio-temporal dynamics over the study region.

Three metrics were used to evaluate the prediction performance: Mean Absolute Error (MAE), Root mean square error (RMSE) (km/h), and Mean Absolute Percentage Error (MAPE). Note the prediction task is performed for each road intersection. The evaluation is done on the basis of intersection. The prediction model outputs predicted speeds at θ cycles after the current signal cycle for all incoming road links that intersect at a road intersection. Prediction results of two selected road intersections are reported in Table I and Table II.

For the two tested intersections and the two time spans, the proposed approach achieves the best prediction accuracies for both three metrics and for three different prediction horizons, with only a few exceptions. During peak hours, our approach outperforms GCN by a larger margin than in non-peak hours, implying that the attentional module play a critical role for predicting peak-hour traffic conditions. The GCN model is superior to the RNN model, which is due to graph convolution operations since RNN only capture temporal correlations.

We also examined the training time of the proposed model.

It takes 115.93 minutes to finish the training whereas the GCN model needs 93.73 minutes. The computational time is dominated by the cost of graph convolution operations. The time costs of spatio-temporal attention modules are relatively minor.

F. Analysis and Discussions

The proposed approach provides a comprehensive solution for traffic speed prediction of signalized road networks. The graph convolution neural network is adopted as the modeling architecture due to its generalization abilities in dynamic traffic modeling. Through spatio-temporal attention modules and graph convolution operations, traffic dependencies and correlations between road links and lanes at both spatial and temporal dimension are effectively captured in the graph neural networks. The network-level macro model provides essential high-level abstracted traffic features that complement the micro traffic features generated from the intersection-level micro model. Without the network-level features, the micro model may fail to characterize traffic dynamics over a larger extent than the examined intersection.

From the raw speed observation dataset, non-linear spatio-temporal correlations and dependencies can be observed, indicating that the travel speed of a signalized road link is affected by its neighboring road links and its historical speed observations. These correlations are subject to changes at different times, thereby posing challenges for short-term real-time prediction. As shown in Table III, the proposed approach can effectively capture these dynamics and maintain small prediction errors for all road links for as long as ten cycles ahead.

TABLE III: Prediction Errors (Observed Minus Predicted Speeds) for Road Links at Intersection #767 During Peak Hours.

Link #	Peak hours			Non-peak hours		
	3	5	10	3	5	10
E49	0.72 ± 1.83	2.09 ± 2.30	3.02 ± 2.43	-6.81 ± 2.32	-6.21 ± 2.01	-5.89 ± 2.96
E42	0.35 ± 2.13	1.66 ± 3.32	4.30 ± 4.27	2.32 ± 2.01	3.84 ± 3.65	4.13 ± 3.65
E15	2.86 ± 3.90	2.78 ± 3.10	1.89 ± 2.90	3.91 ± 0.70	4.54 ± 0.96	4.18 ± 1.37
E12	1.30 ± 2.16	1.91 ± 2.05	-0.71 ± 4.24	-9.56 ± 7.95	-6.81 ± 9.20	-1.81 ± 8.39

IV. CONCLUSIONS

In this study, we have developed and implemented a two-level graph convolutional network model with spatio-temporal attention mechanisms for short-term speed prediction over signalized road networks. Evaluation results demonstrate that the proposed data-driven prediction approach can deliver an effective solution for capturing both network-wide and local spatio-temporal traffic dynamics through two-level network structures, attention mechanisms, and graph convolutional operations. We plan to investigate the applicability of the approach in other cities in the near future. Another possible future research direction may be the integration of other traffic data such as loop detector data to further improve the prediction accuracy.

REFERENCES

- [1] Y. Kamarianakis and P. Prastacos, "Space-time modeling of traffic flow," *Computers & Geosciences*, vol. 31, no. 2, pp. 119–133, 2005.
- [2] Y. Wang and M. Papageorgiou, "Real-time freeway traffic state estimation based on extended kalman filter: a general approach," *Transportation Research Part B: Methodological*, vol. 39, no. 2, pp. 141–167, 2005.
- [3] Y. Yuan, J. Van Lint, R. E. Wilson, F. van Wageningen-Kessels, and S. P. Hoogendoorn, "Real-time lagrangian traffic state estimator for freeways," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 1, pp. 59–70, 2012.
- [4] B. S. Kerner, S. L. Klenov, G. Hermanns, and M. Schreckenberg, "Effect of driver over-acceleration on traffic breakdown in three-phase cellular automaton traffic flow models," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 18, pp. 4083–4105, 2013.
- [5] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Short-term traffic forecasting: Where we are and where we are going," *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 3–19, 2014.
- [6] Y. Li and C. Shahabi, "A brief overview of machine learning methods for short-term traffic forecasting and future directions," *SIGSPATIAL Special*, vol. 10, no. 1, pp. 3–9, 2018.
- [7] Z. Liu, Z. Li, K. Wu, and M. Li, "Urban traffic prediction from mobility data using deep learning," *IEEE Network*, vol. 32, no. 4, pp. 40–46, 2018.
- [8] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: a deep learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865–873, 2015.
- [9] R. Soua, A. Koesdwiady, and F. Karray, "Big-data-generated traffic flow prediction using deep learning and dempster-shafer theory," in *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2016, pp. 3195–3202.
- [10] X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang, and Y. Wang, "Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction," *Sensors*, vol. 17, no. 4, p. 818, 2017.
- [11] Z. Cui, R. Ke, and Y. Wang, "Deep bidirectional and unidirectional lstm recurrent neural network for network-wide traffic speed prediction," *arXiv preprint arXiv:1801.02143*, 2018.
- [12] X. Cheng, R. Zhang, J. Zhou, and W. Xu, "Deeptransport: Learning spatial-temporal dependency for traffic condition forecasting," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–8.
- [13] G. Yang, Y. Wang, H. Yu, Y. Ren, and J. Xie, "Short-term traffic state prediction based on the spatiotemporal features of critical road sections," *Sensors*, vol. 18, no. 7, p. 2287, 2018.
- [14] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *The 3rd International Conference on Learning Representations (ICLR)*, 2014.
- [15] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," in *The 2nd International Conference on Learning Representations (ICLR)*, 2013.
- [16] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *arXiv preprint arXiv:1901.00596*, 2019.
- [17] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.
- [18] X. Feng, J. Guo, B. Qin, T. Liu, and Y. Liu, "Effective deep memory networks for distant supervised relation extraction," in *IJCAI*, 2017, pp. 4002–4008.
- [19] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Advances in neural information processing systems*, 2016, pp. 3844–3852.