# Spatio-Temporal Graph Convolutional Networks for Traffic Forecasting: Spatial Layers First or Temporal Layers First?

Yuen Hoi LAU
yhlauai@connect.ust.hk
Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
Hong Kong, China

Raymond Chi-Wing WONG
raywong@cse.ust.hk
Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
Hong Kong, China

## ABSTRACT

Traffic forecasting is an important and challenging problem for intelligent transportation systems due to the complex spatial dependencies among neighboring roads and changing road conditions in different time periods. Spatio-temporal graph convolutional networks (STGCNs) are usually adopted to forecast traffic features in a road network. Some STGCN models involves spatial layers first and then temporal layers and some other models involves these layers in a reverse order. This creates an interesting research question on whether the ordering of the spatial layers (or temporal layers) first in an existing STGCN model could improve the forecasting performance. To the best of our knowledge, we are the first to study this interesting research problem, which creates a deep insight as a guideline to the research community on how to design STGCN models. We conducted extensive experiments to study a number of representative STCGN models for this research problem. We found that these models with spatial layers constructed before temporal layers has a higher chance to outperform that with temporal layers constructed first, which suggests the future design principle of STGCN models.

## CCS CONCEPTS

• **Applied computing → Transportation**; • **Computing methodologies → Neural networks**.

## KEYWORDS

spatio-temporal graph convolutional networks, traffic forecasting, sequence of modeling, spatial dependencies, temporal dependencies

**ACM Reference Format:**
Yuen Hoi LAU and Raymond Chi-Wing WONG. 2021. Spatio-Temporal Graph Convolutional Networks for Traffic Forecasting: Spatial Layers First or Temporal Layers First?. In *29th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '21), November 2–5, 2021, Beijing, China.* ACM, Beijing, China, 4 pages. https://doi.org/10.1145/3474717.3484207

## 1 INTRODUCTION

Traffic forecasting aims to predict future traffic features including volume, speed, occupancy, demand and travel time of each road segment of a road network. It is useful in many applications such as transportation management, navigation systems, order dispatching and ride sharing. Good traffic forecasting is important in achieving higher efficiency and accuracy of these applications. This task is very challenging due to the complex spatial dependencies among irregular road segments, temporal information from their own and external conditions such as weather and holidays. Spatio-temporal graph neural networks have captured tremendous attention and are usually adopted to accomplish these tasks in recent years thanks to their capability to capture the spatial and temporal dependencies among road segments [20].

There are some assumptions regarding spatio-temporal graph modeling. Road segments are in irregular shapes and regarded as nodes in a graph while traffic sensors gather data on road segments. An adjacent matrix represents the nodes' proximities. A node's future value depends on its own historical values as well as neighboring nodes' information. How to capture spatial and temporal dependencies well is the primary goal. Graph convolution networks (GCN) usually form *spatial layers* while recurrent neural networks (RNN) and temporal convolution neural networks (TCN) become *temporal layers*. Recent studies on most of spatio-temporal graph convolutional networks (STGCNs) are divided into two streams. The first one is to incorporate GCN into RNN [9, 22] and the other is into TCN [21, 19]. GCN is assumed to reflect the spatial dependency relationship among nodes while RNN or TCN entails the temporal information for each node.

Different state-of-the-art STGCNs have different modeling sequence for spatial layers and temporal layers. Some are constructed with spatial layers first [9, 22], some with temporal layers first[21] and a few with both layers gated and fused. Attention mechanisms including additive and scaled dot-product attention are put into some of these models to improve prediction performance. However, none of these studies gives a satisfactory explanation for the modeling sequence of spatial and temporal layers. In view of this, we propose a new research problem: is there a preferable modeling sequence for spatial and temporal layers in STGNN to forecast traffic variables more accurately? To put it succinctly, does a spatio-temporal model constructed with spatial layers first must outperform itself with temporal layers first? To the best of our knowledge, we are the first one to study this research problem.

In this paper, we study this problem by swapping the sequence of the spatial and temporal layers in different STGCNs and comparing their forecasting performance with spatial layers first and

*SIGSPATIAL '21, November 2–5, 2021, Beijing, China*

ACM ISBN 978-1-4503-8664-7/21/11…$15.00
https://doi.org/10.1145/3474717.3484207

temporal layers first respectively for each model by using evaluation metrics including Mean Absolute Errors (MAE) and Root Mean Squared Errors (RMSE). The notable contributions of our paper are summarized as follows:

- To the best of our knowledge, we are the first to explore whether there is a preferable modeling sequence for STGCNs.
- We propose to swap the order of spatial and temporal layers for each type of STGCNs if applicable.
- We conduct experiments on real-world traffic datasets to compare the forecasting performance of each selected state-of-the-art STGCNs with the original modeling sequence and the modified one.

## 2 PROBLEM STATEMENT

Given the road network $G = (V, E, A)$, where each vertex $v \in V$ denotes a road segment and each edge $e \in E$ denotes the proximity between two vertices, an adjacent matrix $A \in \mathbb{R}^{N \times N}$ is derived from the graph and all historical traffic features $X \in \mathbb{R}^{T \times N \times D}$, our problem is to learn and determine whether the composite function $f_T \cdot g_s \cdot h$ representing spatial layers first or $g_S \cdot f_T \cdot h$ representing temporal layers first forecasts $P$ future traffic graph features more accurately given $P'$ historical traffic graph features :

$$[X^{(t-P'+1):t}, G] \xrightarrow{g_T \cdot f_S \cdot h} [X^{(t+1):(t+P)}], \tag{1}$$

$$[X^{(t-P'+1):t}, G] \xrightarrow{f_S \cdot g_T \cdot h} [X^{(t+1):(t+P)}], \tag{2}$$

where $X^{(t-P'+1):t} \in \mathbb{R}^{P' \times N \times D}$ and $X^{(t+1):(t+P)} \in \mathbb{R}^{P \times N \times D}$.

## 3 METHODOLOGY

### 3.1 Overview of Our Proposed Framework

We generalized various state-of-the-art STGCNs with clear modeling sequences into our proposed framework. Figure 1 shows the concept of models constructed with spatial layers first. When the input data are fed into the model, the data processing layers project the lower dimensional input data into higher dimensional traffic features. The layer of GCN captures the spatial dependencies of hidden features. The spatial post-processing layer can be attention mechanisms, diffusion convolutions, residual networks or simply dense layers. These layers form the spatial layers. The layer of TCN or RNN including LSTM and GRU captures the dynamic behaviors of hidden temporal features. Before the output layer, the temporal post-processing layer transforms the hidden features to predicted values. The same concept applies to models constructed with temporal layers first as illustrated in Figure 2. The data processing layers remain unchanged while the sequence of spatial layers and temporal layers is swapped.

### 3.2 Data Processing Layers

The first block of layers are data processing layers, which can be spatial-temporal embedding layers [23], attention mechanisms including Bahdanau attention [1], Luong attention [10] and multi-head scaled dot-product attention mechanism [15], or dense layers, projecting the raw traffic features such as speeds and flows into higher dimensional hidden features.

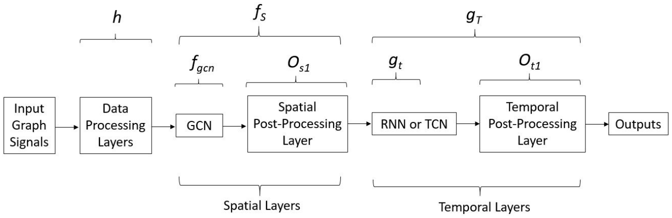

**Figure 1: A concept of models with spatial layers first.**

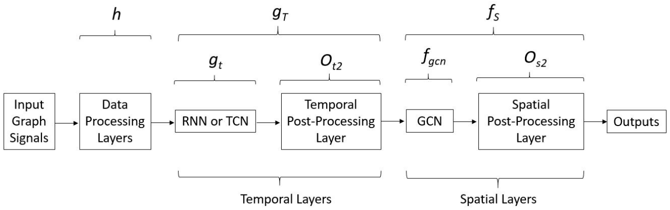

**Figure 2: A concept of models with temporal layers first.**

### 3.3 Spatial Layers

*3.3.1 Graph Convolutional Networks.* GCNs are the building blocks for the spatial layers to learn spatial dependencies among non-Euclidean data. Since there is a multitude of variants of GCNs [16], we generalize the operation of GCNs as follows:

$$h_v^{(i)} = GCN(h_v^{(i-1)},\ f(\{h_u^{(i-1)}, u \in N(v)\})). \tag{3}$$

where $h_v^{(i)}$ denotes the feature vector of node $v$ at the $i^{th}$ layer of GCNs, GCN is an operator to aggregate the information of the node $v$ and its neighbors $N(v)$, $u$ is a neighbor of $v$ and $f$ is a function to capture the information of neighbors.

*3.3.2 Spatial Post-Processing Layer.* After the GCN layer, there is a spatial post-processing layer, which can be attention mechanisms, diffusion convolutions [9], residual networks [18], or dense layers to extract more spatial characteristics. In our modified models, we employ the scaled dot-product attention [15] after the GCN layer to fine-tune spatial dependencies because of its efficiency and comparable performance to other attention mechanisms. The scaled dot-product attention is formulated as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V. \tag{4}$$

This is a spatial self-attention in our modified models with the same hidden features as the query $Q \in \mathbb{R}^{N \times H}$, the key $K \in \mathbb{R}^{N \times H}$ and the value $V \in \mathbb{R}^{N \times H}$. $d_k$ is the number of hidden features. $N$ is the number of nodes and $H$ is the number of hidden features.

### 3.4 Temporal Layers

The building blocks of temporal layers are Recurrent Neural Networks (RNNs) or Temporal Convolutional Networks (TCNs) [3]. RNNs including GRU [4] and LSTM [6] have shown the capability of modeling the temporal dependency to a large extent. GRU is preferred over LSTM because of its efficiency and comparable forecasting performance. TCNs are also widely adopted due to its efficiency. We can stack multiple GRU layers and employ the sequence to sequence architecture for multiple step ahead forecasting.

*3.4.1 Temporal Convolutional Networks.* Compared to RNNs, TCNs have the advantages of being able to handle long-range sequences in a non-recursive manner, enabling parallel computation and alleviating the gradient explosion problem. Let $F = (f_1, f_2, ..., f_K)$ be filters, $H^{(t)} \in \mathbb{R}^{P' \times F}$ be hidden features from previous layers, and $h^{(t)} \in \mathbb{R}^F$ be a node in $H^{(t)}$. The operation of TCNs at time $t$ is defined as

$$F_{*d} H^{(t)} = \sum_{k=1}^{K} f_k h^{(t)}_{P'-(K-k)\times d} \tag{5}$$

where $d$ is the dilation factor which controls the skipping distance. Stacking multiple TCNs can capture the information of longer range sequences.

*3.4.2 Temporal Post-Processing Layer.* The layer after RNN or TCN layers can be attention mechanisms, gated linear unit [21, 18] or fully connected layers to capture temporal dependencies of hidden features. In our modified models, we employ the same self-attention mechanism as that in spatial layers.

# 4 EXPERIMENTS

We set the scope of our research on the models which take only traffic features including speeds and/or flows as input. It is shown by many papers that additional data such as weather and holidays can enrich traffic conditions to improve the forecasting capability of STGCNs. Therefore, we would like to focus on finding out a better architecture for STGCNs which have a clear spatio-temporal modeling sequence without these additional data. Whether a model is influential or not is our major consideration for being selected in our experiments. There are a number of papers [9, 5, 2, 22, 12, 17] proposing spatial-layer-first models. We chose **DCRNN** [9], **ASTGCN** [5] and **T-GCN** [22] as representative models in our experiments for comparison since they had at least 192 citations within 3 years but other models did not. We also selected **AGCRN** [2] due to an influential work regarding a similar idea of the adaptive graph convolution [8]. There are a few papers [21, 18, 7] proposing temporal-layer-first models. We chose **STGCN** [21] and **Graph WaveNet** [18] as representative models in our experiments for comparison since they had at least 144 citations within 3 years but LSGCN [7] did not.

It is worth mentioning that hybrid models proposed in [13, 14, 23, 11] are not under our framework considering spatial/temporal layers first, followed by other types of layers, since hybrid models consider both spatial and temporal layers at the same time in the form of a graph including spatial and temporal nodes. In [13], it is possible that the hybrid model USTGCN could perform worse than Graph WaveNet which follows our framework in PeMSD4 datasets. In [14], the hybrid model STSGCN could perform worse than DCRNN which follows our framework in PeMS07 datasets. In [23], the hybrid model GMAN could perform worse than DCRNN and Graph WaveNet in Xiamen and PEMS datasets. However, it is interesting to explore what exact forms of graphs in hybrid models could perform better, which could be regarded as a future work.

By swapping the spatial layers and temporal layers and compare the forecasting performance of the different modeling sequence for each of the following models.

**Table 1: Performance Comparison of models with spatial layers first and temporal layers first using original datasets**

| Models | 15 min | | 30 min | | 60 min | |
|---|---|---|---|---|---|---|
| | MAE | RMSE | MAE | RMSE | MAE | RMSE |
| DCRNN-s (original) | 15.71 | 29.12 | 16.10 | 29.34 | 16.67 | 29.70 |
| DCRNN-t (ours) | 15.89 | 28.90 | 16.51 | 29.42 | 17.44 | 30.10 |
| DCRNN-s-att (ours) | **14.96** | **27.79** | **15.40** | **28.07** | **16.07** | **28.52** |
| DCRNN-t-att (ours) | 20.71 | 32.96 | 20.68 | 32.93 | 20.68 | 32.93 |
| ASTGCN-s (original) | **20.2** | **31.58** | **22.24** | **34.46** | **26.75** | **40.64** |
| ASTGCN-t (ours) | 20.36 | 31.86 | 22.55 | 35 | 27.38 | 41.55 |
| T-GCN-s (original) | **3.30** | 5.26 | 3.82 | 6.27 | 4.63 | 7.59 |
| T-GCN-t (ours) | 3.66 | 5.49 | 3.99 | 6.29 | 4.80 | 7.47 |
| T-GCN-s-att (ours) | 3.33 | **5.25** | **3.82** | **6.27** | **4.62** | **7.59** |
| T-GCN-t-att (ours) | 3.42 | 5.42 | 3.88 | 6.30 | 4.60 | 7.64 |
| AGCRN-s (original) | 18.96 | 31.10 | 19.72 | 32.45 | 21.28 | 35.13 |
| AGCRN-t (ours) | 18.75 | **30.31** | **19.51** | **31.62** | 21.19 | **34.19** |
| AGCRN-s-att (ours) | **18.72** | 30.64 | **19.51** | 32.22 | **20.88** | 34.67 |
| AGCRN-t-att (ours) | 20.06 | 31.63 | 21.66 | 34.07 | 25.37 | 39.22 |
| STGCN-s (ours) | 2.73 | 5.06 | 3.93 | 7.52 | 3.61 | 6.84 |
| STGCN-t (original) | 2.74 | 5.06 | 3.89 | 7.32 | 3.66 | 6.86 |
| STGCN-s-att (ours) | **2.71** | **4.99** | 3.45 | 6.29 | **3.40** | **6.39** |
| STGCN-t-att (ours) | 2.71 | 4.99 | **3.42** | **6.18** | 3.45 | 6.41 |
| GraphWaveNet-s (ours) | **2.85** | **5.34** | **3.35** | **6.54** | **4.01** | **8.06** |
| GraphWaveNet-t (original) | 2.99 | 5.74 | 3.62 | 7.20 | 4.55 | 9.06 |
| GraphWaveNet-s-att (ours) | 3.12 | 5.82 | 3.73 | 7.27 | 4.57 | 9.00 |
| GraphWaveNet-t-att (ours) | 3.05 | 5.82 | 3.69 | 7.28 | 4.60 | 9.03 |

## 4.1 Experimental Settings

We keep the original parameter settings for each model and use the original datasets employed by each corresponding paper for all experiments. We adopt two common metrics to measure the forecasting performance of different models, including Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

We employed original datasets for each model. We collected four sets of data from GitHub released by authors of each selected paper. All the datasets are aggregated every 5 minutes. Thus, each model is tested with its original dataset in its original paper for the best tuned parameters. (1) **METR-LA** [9] is used to test DCRNN and Graph WaveNet. (2) **PeMSD4** [5] is employed to test AGCRN and AGCRN. (3) **Los-loop** [22] is employed to test T-GCN. (4) A medium dataset **PeMSD7(M)** [21] is employed to test STGCN.

## 4.2 Forecasting Performance Comparison

The forecasting performance for each model is shown in Table 1. The model name appended with suffix "s" is the model constructed with spatial layers first and appended with suffix "t" is with temporal layers first. The suffix "att" means that a single head scaled dot-product attention is added to spatial layers or to temporal layers to capture more spatio-temporal dependencies.

DCRNNs constructed with spatial layers first has better performance than those constructed with temporal layers first in terms of MAE and RMSE in the prediction of traffic features in future 15th, 30th and 60th minutes. With the spatial layers constructed first and the attention layer added, DCRNN-s-att performs better than DCRNN-s in almost all aspects while attention layers cannot make DCRNN-t-att perform better. ASTGCN is built with both spatial and temporal attention mechanisms before the GCN and TCN layers. We do not further add attention mechanisms to it in our experiments because of the existence of attention mechanisms in the model. With the GCN layer before the TCN layer, ASTGCN-s outperforms its alternative for all prediction tasks. T-GCN has a clear and simple modeling sequence. GCN is the spatial layer while GRU is the temporal layer. The prediction results of T-GCN-s-att are the best. It may imply that with spatial layers constructed before temporal layers and with attention added, the model is likely to outperform its alternatives. AGCRN-s-att performs the best in terms of MAE while AGCRN-t performs the best regarding RMSE as the evaluation metric for all the forecasting tasks. The attention mechanisms improve the results of AGCRN-s while they make AGCRN-t perform worse. STGCN-s-att outperforms other variants except the future 30th-minute forecasting. With spatial layers constructed first, STGCN-s has lower prediction errors than STGCN-t except forecasting 30th-minute traffic features. Graph WaveNet-s has the best forecasting performance. Graph WaveNet-s-att performs worse than Graph WaveNet-s, meaning that attention layers disturb the forecasting process. WaveNet-s-att performs a bit better than WaveNet-s-att in four out of six results. These results also imply that with spatial constructed first, the forecasting results are likely to be better.

### 4.3 Discussion of Results

There are two effects influencing the forecasting performance of the models and they are originated from the characteristics of datasets: spatial dependencies among neighbor nodes and temporal dependencies for each node for different time periods. A possible reason for the observed results is that the effect of spatial dependencies outweighs that of temporal ones in those datasets. For the STGCNs constructed with spatial layers first, hidden features of each node have incorporated those of neighbors after the GCN operation at each time period. If datasets have strong spatial dependencies, spatial layers can generate meaningful hidden features among neighbor nodes to reflect more accurate traffic conditions. Since it is very likely that the traffic of neighbor nodes at the current timestamp may affect the traffic of the current node at the next timestamp, considering the spatial layers first somehow already captures some temporal information, which could explain why using the spatial layers first could improve the performance. Afterwards, temporal layers capture the temporal dependencies of those meaningful hidden features from spatial layers for each node and make predictions, which could improve the prediction performance. Another case is that spatial layers may not generate meaningful hidden features for each node if spatial dependencies are not strong enough. Afterwards, temporal layers will process less meaningful hidden features and predictions are made less accurately. If the effect of temporal dependencies is stronger than that of spatial ones, STGCNs constructed with temporal layers first may generate more meaningful hidden features and more accurate predictions.

## 5 CONCLUSION

In this paper, we proposed a new problem for the field of spatio-temporal graph convolutional networks, which states that whether the modeling sequence for spatial layers and temporal layers matters. We tried to give an answer by swapping the sequence of spatial and temporal layers for each of the selected models which have clear modeling sequences. By conducting experiments, for most of the selected models, if constructed with spatial layers before temporal layers and with attention added, they have a higher chance of beating their alternatives. This indicates that STGCNs constructed with spatial layers first may have an advantage over that with temporal layers first. A possible reason is that the effect of spatial dependencies outweighs that of temporal ones in those datasets, leading to a relative advantage in models constructed with spatial layers first. More experiments and mathematical proof help to explore how the modeling sequence of STGCNs matters.

## REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *ICLR*. 2015.

[2] Lei Bai et al. "Adaptive Graph Convolutional Recurrent Network for Traffic Forecasting". In: *NIPS*. 2020.

[3] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. *An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling*. 2018. arXiv: 1803.01271 [cs.LG].

[4] Junyoung Chung et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling". In: *arXiv preprint arXiv:1412.3555* (2014).

[5] Shengnan Guo et al. "Attention based spatial-temporal graph convolutional networks for traffic flow forecasting". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 922–929.

[6] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[7] Rongzhou Huang et al. "LSGCN: Long Short-Term Traffic Prediction with Graph Convolutional Networks". In: *IJCAI*. 2020, pp. 2355–2361.

[8] Ruoyu Li et al. "Adaptive graph convolutional neural networks". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.

[9] Yaguang Li et al. "Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting". In: *International Conference on Learning Representations (ICLR '18)*. 2018.

[10] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. "Effective approaches to attention-based neural machine translation". In: *EMNLP*. 2015.

[11] Li Mengzhang and Zhu Zhanxing. "Spatial-Temporal Fusion Graph Neural Networks for Traffic Flow Forecasting". In: *Proceedings of the AAAI conference on artificial intelligence*. 2021.

[12] Cheonbok Park et al. "ST-GRAT: A Novel Spatio-temporal Graph Attention Networks for Accurately Forecasting Dynamically Changing Road Speed". In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2020, pp. 1215–1224.

[13] Amit Roy et al. "Unified Spatio-Temporal Modeling for Traffic Forecasting using Graph Neural Network". In: *IJCNN*. 2021.

[14] C. Song et al. "Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 2020, 914–921.

[15] Ashish Vaswani et al. "Attention is all you need". In: *NIPS*. 2017, 5998–6008.

[16] Zonghan Wu et al. "A Comprehensive Survey on Graph Neural Networks". In: *IEEE Transactions on Neural Networks and Learning Systems* 32.1 (2021), pp. 4–24.

[17] Zonghan Wu et al. "Connecting the dots: Multivariate time series forecasting with graph neural networks". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 753–763.

[18] Zonghan Wu et al. "Graph WaveNet for Deep Spatial-Temporal Graph Modeling". In: *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*. 2019.

[19] Sijie Yan, Yuanjun Xiong, and Dahua Lin. "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition". In: *AAAI*. 2018.

[20] Xueyan Yin et al. *A Comprehensive Survey on Traffic Prediction*. 2021. arXiv: 2004.08555v1 [eess.SP].

[21] Bing Yu, Haoteng Yin, and Zhanxing Zhu. "Spatio-temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting". In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*. 2018.

[22] Ling Zhao et al. "T-GCN: A Temporal Graph Convolutional Network for Traffic Prediction". In: *IEEE Transactions on Intelligent Transportation Systems* 21.9 (2020), 3848–3858. ISSN: 1558-0016.

[23] Chuanpan Zheng et al. "GMAN: A Graph Multi-Attention Network for Traffic Prediction". In: *AAAI*. 2020, pp. 1234–1241.