2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA)

DASTGCN: Dynamic Attention Based Spatio-Temporal Graph Convolutional Network for Traffic Forecasting

Peng Liu[1]

[1]College of Information Engineering, Shanghai Maritime University, Shanghai, 201306, China
liupeng0200@stu.shmtu.edu.cn

***Abstract*—Traffic flow prediction plays a very important role in transportation, and accurate traffic flow prediction can bring great convenience to transportation. There are two main limitations in the existing traffic flow prediction models: First, most models do not separate the dynamic and periodic spatio-temporal features, resulting in models that do not fully exploit the spatio-temporal features of traffic flow. Second, most models perform poorly in extracting temporal features, and it is difficult to aggregate node information for longer time steps. Therefore, this paper proposes a new Dynamic Attention Based Spatio-Temporal Graph Convolutional Network(DASTGCN). We design a spatio-temporal dynamic attention module and a multi-scale gated convolution to solve the above problem. Experiments on two real datasets show that our DASTGCN model performs better than other models.**

***Keywords*—traffic forecasting, deep learning, attention mechanism, graph convolution**

I. Introduction

In recent years, many countries are committed to the development of Intelligent Transportation Systems (ITS) [1] for efficient traffic management. Traffic flow prediction is an indispensable part of the intelligent transportation system, especially for the highway with large traffic volume and fast driving speed, due to the relative closure of the highway, once the congestion occurs, it will seriously affect the capacity.

Although most of the models show good performance in traffic flow prediction, there are still the following problems to be solved in the existing methods. First, the traffic flow is affected by the travel time period and holidays, as shown in Figure 1(A), the difference between the traffic flow of the same node on weekdays (blue) and rest days (orange) is very obvious. The location relationship between nodes has a great impact on the variation of traffic flow, as shown in Figure 1(B), the correlation between nodes A and B is stronger in the morning and becomes weaker in the evening. On the contrary, the correlation between nodes A and E is weaker in the morning and stronger in the evening.Most models use a static approach to model node correlations, which severely limits the ability to learn dynamic traffic patterns. Second, graph convolutional networks perform poorly in temporal feature extraction, and most GCN-based models suffer from over-smoothing, making it difficult to capture the spatio-temporal dependence of long distances.

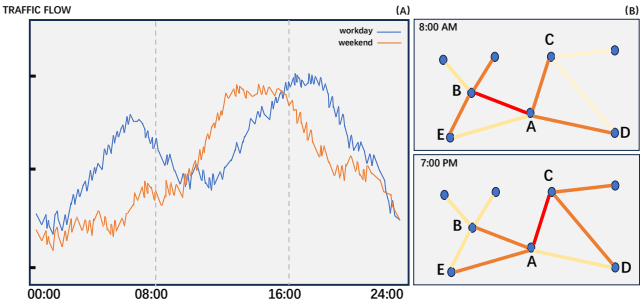

Figure 1. The spatial-temporal correlation diagram of traffic flow.

In order to solve the above problems, the main work in this paper is as follows:

- We designed a dual backbone graph neural network model(DASTGCN) based on dynamic attention[2]. The dual backbone network solves the problem that the model can extract both regular spatio-temporal correlations and capture dynamic spatio-temporal correlations in the traffic flow.
- A new spatio-temporal correlation matrix is used to replace the static matrix[3], the spatio-temporal dynamic attention module is further designed to extract complex dynamic spatio-temporal features in the traffic flow by dynamically adjusting the attention weight matrix. In addition, multi-scale temporal gated convolution[4] replaces the traditional temporal convolution to extract node features with longer time steps.
- Experiments have been carried out on real motorway traffic datasets PEMS04, PEMS08 which have show our model get better prediction performance compared to other models.

II. Related Work

A. Problem definitions

We define a traffic network as an undirected graph G = (V, E, A), where V is a finite set of $|V| = N$; E is the set of edges

979-8-3315-0976-7/25/$31.00 ©2025 IEEE

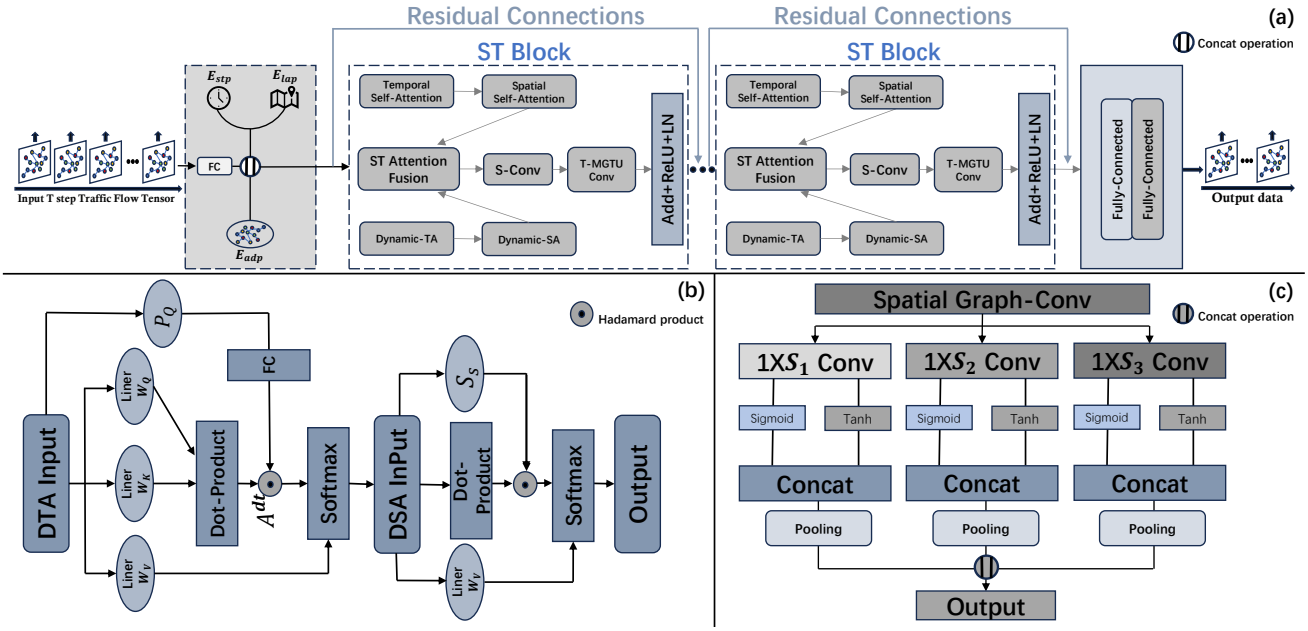

Figure 2. The framework of Dynamic Attention Based Spatio-Temporal Graph Convolutional Network(DASTGCN).

denoting the connectivity between nodes; and A denotes the adjacency matrix of the graph G. F is the feature of each node in the graph, here it represents the traffic flow.

The model uses past traffic flow data to predict future data, the prediction problem can be expressed by the following equation:

$$[X_{t-T+1}, \ldots, X_t] \xrightarrow{F(\cdot)} [X_{t+1}, \ldots, X_{t+T'}] \tag{1}$$

where each frame $X_i \in \mathbb{R}^{N \times d}$, N is the number of spatial nodes. d is the dimension of the input feature which equals 1 in our case, standing for traffic volume.

III. Methodology

The architecture of our model is shown in Figure 2, which consists of the spatio-temporal embedding layer, the spatio-temporal block consisting of the spatio-temporal attention layer and the spatio-temporal convolution layer, and the output layer. Then, we will introduce the role and working principle of each layer.

A. Spatio-temporal Embedding Layer

In order to make the data have richer features and facilitate the extraction of features contained in the data, we design a spatial-temporal embedding layer to incorporate the necessary knowledge into the model[5], including spatio-temporal position embedding, graph Laplace embedding[6] and graph adaptive embedding[7]. The original data input X is transformed through the fully connected layer into $X_{\text{data}} \in \mathbb{R}^{N \times F \times T}$, N is the number of nodes, F is the embedding dimension and T is the time step.

We use the graph Laplace matrix, which describes the positional relationships between nodes in the graph.We obtain a standardised Laplace matrix as follows:$\widehat{A} = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$, where A is the adjacency matrix, D is the degree matrix and I is the unit matrix. Then,the eigenvalue matrix Λ and eigenvector matrix U are obtained by eigenvalue decomposition. Generating spatial graph Laplace embedding $E_{\text{lap}} \in \mathbb{R}^{N \times d_{\text{lap}}}$ using linear projections of the k smallest non-trivial feature vectors, where d_{lap} is the Laplace embedding dimension. Spatial graph Laplacian embedding effectively models the road network and preserves its topological structure.

The relationship between traffic flows between nodes is not only determined by the position, but is also influenced by the time of day, spatio-temporal positional coding can effectively represent the relative positional relationships between nodes at each time step.$E_{\text{stp}} \in \mathbb{R}^{T \times N \times d_{\text{stp}}}$ obtained by sin/cos positional encoding, where d_{stp} is the embedding dimension.

It is easy to infer that the traffic flow of a node at a certain time is often more similar to the traffic flow at nearby times. In addition, we also know that different nodes often have completely different traffic patterns. We designed a graph adaptive embedding $E_{\text{adp}} \in \mathbb{R}^{N \times d_{\text{adp}}}$ to adaptively complements the complex spatio-temporal relationships in traffic flow and E_{adp} shares parameters across all time[8].

The above three embeddings are concatenated with the original data to get the output X_{emb} as follows:

$$X_{\text{emb}} = \text{Concat}(X_{\text{data}}, E_{\text{adp}}, E_{\text{lap}}, E_{\text{stp}}) \tag{2}$$

B. Spatio-temporal Attention Layer

The spatio-temporal attention layer consists of the spatio-temporal self-attention layer and the spatio-temporal dynamic attention layer[9]. The self-attention layer extracts periodic and regular spatio-temporal features in traffic flow, while the dynamic attention layer extracts dynamic and complex spatio-temporal features[10][11].

1. Spatio-temporal Self-attention Layer

We capture the inherent temporal correlation in traffic flow through the temporal self-attention[12]. Formally, for the node N, we first obtain the Q^{te}, $K^{te}$and V^{te}matrices:

$$Q^{te} = X_{\text{emb}} W_Q^{te}, K^{te} = X_{\text{emb}} W_K^{te}, V^{te} = X_{\text{emb}} W_V^{te} \quad (3)$$

where W_Q^{te},W_K^{te},W_V^{te} are learnable parameters. Then we calculate the self-attention score as:

$$A^{te} = \text{Softmax}\left(\frac{Q^{te} K^{te^\top}}{\sqrt{d_{te}}}\right), A^{te} \in \mathbb{R}^{T \times T} \quad (4)$$

We obtain the output of temporal self-attention $Z^{TE} \in \mathbb{R}^{N \times T \times d_{te}}$ as:

$$Z^{TE} = \text{TSA}(Q^{te}, K^{te}, V^{te}) = \text{Softmax}(A^{te}) V^{te} \quad (5)$$

Similarly, the spatial self-attention layer performs as:

$$Q^{sp} = Z^{TE} W_Q^{sp}, K^{sp} = Z^{TE} W_K^{sp}, V^{sp} = Z^{TE} W_V^{sp} \quad (6)$$

Finally, we leverage the output of the spatio-temporal self-attention layers:

$$Z^{ST} = \text{SSA}(Q^{sp}, K^{sp}, V^{sp}) = \text{Softmax}(A^{sp}) V^{sp} \quad (7)$$

2. Spatio-temporal Dynamic Attention Layer

Traffic flow is extremely dynamic and it is difficult for self-attention to fully extract spatio-temporal features, so we need to dynamically adjust the attention module to capture the dynamic spatio-temporal features of traffic flow[13].

We use deformable attention to extract temporal features, and the architecture of this module is shown in Figure 2(b). We first get Q^{te} while we input a light quantum network θ_{offset} to get the offset $\Delta\theta_{\text{offset}}(Q)$. Offsetting the nodes in the input X_{emb} according to the offset, calculated K^{te},V^{te} at the offset position P_Q. The attention score A_{dt} is obtained by X_{emb} and X_{emb} making a Hadamard product with the sum after full connectivity. The attentional score A^{dt} is obtained by making a Hadamard product of P_Q, with Q^{te} and K^{te}, after full connectivity. The output $Z^{DT} \in \mathbb{R}^{N \times T \times d_{dt}}$ is obtained from the attention scores A^{dt} and V^{te}, which can be calculated by:

$$Z^{DT} = \text{DTA}(Z_Q, P_Q, Z^{ST}) = \text{Softmax}(A^{dt}) V^{te} \quad (8)$$

We use the graph convolution network combined with the attention to extract dynamic spatial features of traffic flow[14]. Calculate the Q^{sp}, K^{sp} and V^{sp} of the Z^{DT} in order to obtain the attention score A^{dsp}. Then, we calculate Spatial correlation weight matrix S_t, its calculation is similar to that of A^{dsp}, which is obtained from the following equation:

$$S_t = \text{Softmax}\left(\frac{Z_t^{DT} Z_t^{DT^\top}}{\sqrt{d_{dt}}}\right), S_t \in \mathbb{R}^{N \times N} \quad (9)$$

The spatial correlation weight matrix S_t is used to adjust attention score matrix A^{dsp} to obtain the dynamic weight matrix, The final output $Z^{DST} \in \mathbb{R}^{N \times T \times d_{dst}}$ is obtained and can be expressed as follows:

$$Z^{DST} = (Q^{sp}, K^{sp}, V^{sp}) = \text{Softmax}(A^{dsp} \odot S_t) V^{sp} \quad (10)$$

Fusing the outputs of the two attentions Z^{ST} and Z^{DST}. Eventually, the output of the spatio-temporal attention layer is obtained as:

$$Z^{\text{out}} = \text{add}(Z^{ST}, Z^{DST}) \quad (11)$$

C. Spatio-temporal Convolution Layer

The output Z^{out} of the spatio-temporal attention layer goes to the spatio-temporal convolution layer, where spatial features of traffic flow are captured by spatial graph convolution[15] and temporal features are captured by multi-scale gated temporal convolution[16].

Spatio Convolution Layer

We consider the daily traffic flow detected by a particular detector as a vector (recorded every five minutes, 288 records in a day d^t=288), thus the multi-day traffic flow can be viewed as a sequence of vectors[9]: $X_n^f = (W_{n1}, W_{n2}, W_{n3}, \ldots, W_{nD}),\ W_{nD} \in \mathbb{R}^{d_t}$.

The distribution probability of the node n can be expressed as:

$$P_n(X_d = m_{nd}), m_{nd} = \frac{\|W_{nd}\|_2}{\sum_{d=1}^{D} \|W_{nd}\|_2} \quad (12)$$

We use the Wasserstein distance to measure the difference in probability distributions and get:

$$W[u, v] = \inf_{\gamma \in \pi[u,v]} \iint_{(x,y)} \gamma(x, y) d(x, y)\, dx\, dy \quad (13)$$

where γ is the joint probability distribution, the edge distribution is u,v and inf denotes the lower bound. Finally, the similarity matrix A_{stag} is obtained using the cosine distance as a cost function.

We use the degree matrix D and the adjacency matrix A representing the node relationship to obtain the algebraic representation of the Laplacian matrix L = D - A, We replace the adjacency matrix A in the equation with our spatio-temporal correlation graph A_{stag}, having L = D - A_{stag}. Its normalised form is $L = I_N - D^{-1/2} A_{\text{stag}} D^{-1/2},\ L \in \mathbb{R}^{N \times N}$.

Then, we perform the eigenvalue decomposition to obtain the eigenvalue matrix Λ and eigenvector matrix U. Taking the traffic flow at moment t as an example, all the nodes on the graph are $X = X_t^f \in \mathbb{R}^N$, and the graph Fourier transform is defined as $\hat{X} = U^\top X$. Next, we perform spatial convolution operation. Nodes X on the graph G is filtered by the kernel function g_θ: $g_\theta * GX = g_\theta(L)X = g_\theta(U \Lambda U^\top)X = U g_\theta(\hat{\Lambda}) U^\top X$, where $*G$ denotes the graph convolution operation.

However, when the traffic network is large, the computational cost of L is high. Therefore, the Chebyshev polynomial approximation (Simonovsky and Komodakis 2017) is used to solve this problem effectively:

$$g_\theta * GX = g_\theta(L)X = \sum_{k=0}^{K-1} \theta_k T_k(\tilde{L}) X \quad (14)$$

where the parameter $\theta \in \mathbb{R}^k$ is a vector of polynomial coefficients, $\tilde{L} = \frac{2}{\lambda_{\max}} L - I_N$, $\lambda_{\max}$ is the maximum eigenvalue of the Laplace matrix.

We can think of g as a function $g_\theta(L)$ about the Laplace matrix L, each convolution is equivalent to passing information about neighbouring nodes.The recursive definition of Chebyshev polynomials is:

$$T_k(X) = 2XT_{k-1}(X) - T_{k-2}(X) \tag{15}$$

Here $T_0(X) = 1,\ T_1(X) = X$, we use an approximate expansion of the Chebyshev polynomials to solve for information about the surrounding neighbours of order 0 to k-1 centred on each node in this graph. Using $\mathrm{ReLU}(g_\theta * GX)$ as the final activation function, the output is $Z^l \in \mathbb{R}^{N\times T\times d_l}$.

1. Multi-scale Gated Temporal Convolution

We use the M-GTC Layer to capture temporal features of traffic flow data.The specific structure is shown in Figure 2(c), consisting of three Gated Tanh Units (GTUs)[17], the module has 3 different receptive fields.

Previous GTUs doubled the number of channels by using convolution, the convolution kernel is denoted as $\Gamma \in \mathbb{R}^{1\times S\times c^l\times 2c^l}$, its size is 1 × S. We superimpose multiple GTUs with expanding receptive fields, by doing so, the ability of the model to capture temporal features at long time steps is improved[18]. Thus, obtaining our multi-scale gated temporal convolution (M-GTC):

$$\begin{aligned} Z^l_{\text{out}} = \text{M-GTC}(Z^l) = \mathrm{ReLU}(\mathrm{Concat}(\mathrm{Pooling}(\Gamma_1 * Z^l), \\ \mathrm{Pooling}(\Gamma_2 * Z^l), \mathrm{Pooling}(\Gamma_3 * Z^l)) + Z^l) \end{aligned} \tag{16}$$

where $\Gamma_1, \Gamma_2, \Gamma_3$ denote convolution kernels, sizes are $1 \times S_1,\ 1 \times S_2,\ 1 \times S_3$.Concat denotes the splicing operation and Pooling denotes the pooling operation, ReLU is the activation function. Eventually, the output is $Z \in \mathbb{R}^{N\times T\times d_{\text{st}}}$.

D. Output Layer

The output of the spatio-temporal convolution layer goes into multiple fully connected layers, the final fully connected layer uses ReLU as the activation function. Then using 2d convolution,we obtain the prediction $\hat{Y}$ of our DASTGCN model, the formula is expressed as follows:

$$\hat{Y} = \mathrm{conv}\left(\mathrm{FC}\left(\mathrm{ReLU}(Z)\right)\right) \tag{17}$$

IV. Experiments

A. Datasets

We conducted experiments on two California highway traffic datasets, PEMSD4 and PEMSD8. The datasets were collected by the California Transportation Performance Measurement System (PeMS) [19]. Traffic data were aggregated by highway sensors at 5-minute intervals. This dataset includes the traffic flow of highways in various cities in California. There are about 4,000 collector nodes that collect traffic flow on highways in real time.

We removed some nodes that were too close to each other to ensure that each node detected meaningful traffic flow. Finally, PEMSD4 has 307 detectors and PEMSD8 has 170 detectors. Each node detects traffic flow once every 5 minutes, which means 288 traffic data per day, and linear interpolation is used to fill missing values. In addition, the data were transformed by zero-mean normalisation x = x - mean(x) so that the mean was 0. The training, validation and test set ratios were divided into 6:2:2.

We use two widely used traffic prediction metrics, i.e., MAE, RMSE to evaluate the performance of our models. Based on previous work, we chose the average performance of the 12 time-step predictions of the model predictions PEMSD04, PEMSD08 as a comparative value for the performance of each model.

B. Baseline Model

The comparison results are shown in the Table 1 below, and it can be seen that our model has a good performance compared to most of the models on the two datasets PEMSD04, PEMSD08.

Table 1: Performance on PEMSD04, PEMSD08.

Dataset	PEMSD04		PEMSD08	
Model	**RMSE**	**MAE**	**RMSE**	**MAE**
VAR	51.73	33.76	31.21	21.41
LSTM	45.82	29.45	36.96	23.18
GRU	45.11	28.65	35.95	22.20
STGCN	38.27	25.15	27.87	18.88
GLU-STGCN	38.41	27.28	30.78	20.99
GeoMAN	37.84	23.64	28.91	17.84
MSTGCN	35.65	22.73	26.47	17.47
ASTGCN	32.82	21.80	25.27	16.63
DCRNN	36.25	23.54	28.28	18.14
STSGCN[20]	33.58	20.98	25.64	16.35
STGODE[21]	32.82	20.85	26.25	16.82
GTS	32.95	20.96	26.08	16.49
DASTGCN	**31.21**	**19.26**	**24.56**	**15.62**

C. Ablation Study

To further investigate the validity of the different parts in DASTGCN, we compared DASTGCN with the following variants. RemM-GTC: Removed M-GTC and temporal features are extracted by temporal convolution. RemD-STA: Removed Spatio-temporal Dynamic Attention Layer. RemSTA: Removed Spatio-temporal Self-attention Layer. We did

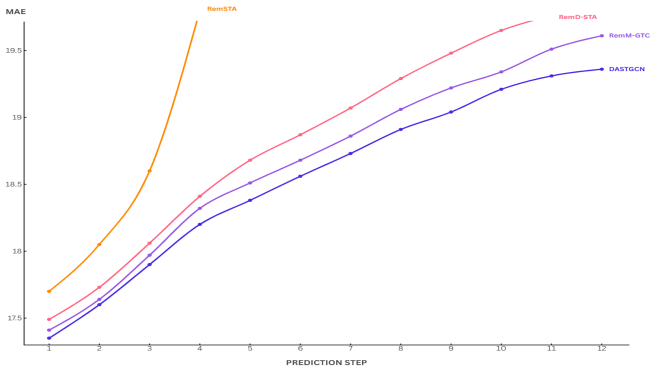

Figure 3. Ablation Study on PEMS04.

a comparison test on PEMSD04 and the results are shown in Figure 3, where we can see that the spatio-temporal self-attention module is crucial in DASTGCN. RemD-STA shows

that both regular and dynamic spatio-temporal features are important for flow prediction. In addition, M-GTC performs well in capturing spatio-temporal features at long time steps.

V. Conclusion

In this work, we propose a new attention-based spatio-temporal graph convolutional model DASTGCN. Specifically, we use spatio-temporal self-attention layer and spatio-temporal dynamic attention layer to better capture the spatio-temporal correlation of traffic flows. We use spatio-temporal convolution combined with a multi-scale gating mechanism to better extract spatio-temporal features. The use of a dual backbone network provided some inspiration for the development of subsequent models. After experiments on two datasets and comparison with other models, it is proved that our model performs well. In addition, the ablation experiment also proves that our model architecture is reasonable and interpretable. In future research, we may conduct more research on data embedding and improve the architecture of the dual backbone network.

References

[1] Nour Eddin El Faouzi, Henry Leung, and Ajeesh Kurian. "Data fusion in intelligent transportation systems: Progress and challenges – A survey". In: *Information Fusion* 12.1 (2012), pp. 4–10.

[2] Xia Dawen et al. "Attention-based spatial–temporal adaptive dual-graph convolutional network for traffic flow forecasting". In: *Neural computing applications* (2023).

[3] Liangzhe Han et al. "Dynamic and Multi-faceted Spatio-temporal Deep Learning for Traffic Speed Forecasting". In: *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2021.

[4] Xu Chen et al. "TSSRGCN: Temporal Spectral Spatial Retrieval Graph Convolutional Network for Traffic Flow Forecasting". In: (2020).

[5] Yuchen Fang et al. "Spatio-Temporal meets Wavelet: Disentangled Traffic Flow Forecasting via Efficient Spectral Graph Attention Network". In: (2021).

[6] Jiawei Jiang et al. "Pdformer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 37. 4. 2023, pp. 4365–4373.

[7] Hangchen Liu et al. "STAEformer: spatio-temporal adaptive embedding makes vanilla transformer SOTA for traffic forecasting". In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM), Birmingham, UK*. 2023, pp. 21–25.

[8] X. Zhang et al. "AdpSTGCN: Adaptive spatial–temporal graph convolutional network for traffic forecasting". In: *Knowledge-based systems* Oct.9 (2024), p. 301.

[9] Shiyong Lan et al. "Dstagnn: Dynamic spatial-temporal aware graph neural network for traffic flow forecasting". In: *International conference on machine learning*. PMLR. 2022, pp. 11906–11917.

[10] Yanshen Sun, Kaiqun Fu, and Chang Tien Lu. "DG-Trans: Dual-level Graph Transformer for Spatiotemporal Incident Impact Prediction on Traffic Networks". In: (2023).

[11] Liu Shang et al. "Double Branch Spatial-temporal Graph Convolutional Neural Network for Traffic Flow Prediction". In: *Information and Control* 52.3 (2023), pp. 391–404, 416.

[12] Zhihong Chang, Chunsheng Liu, and Jianmin Jia. "STA-GCN: Spatial-Temporal Self-Attention Graph Convolutional Networks for Traffic-Flow Prediction". In: *Applied Sciences (2076-3417)* 13.11 (2023).

[13] Zequan Li et al. "Dynamic spatial aware graph transformer for spatiotemporal traffic flow forecasting". In: *Knowledge-based systems* Aug.3 (2024), p. 297.

[14] Hong Zhang et al. "Research on Traffic Flow Forecasting Based on Dynamic Spatial-Temporal Transformer:" in: *Transportation Research Record* 2678.7 (2024), pp. 301–313.

[15] Xiyue Zhang et al. "Traffic Flow Forecasting with Spatial-Temporal Graph Diffusion Network". In: (2021).

[16] Lei Bai et al. "Adaptive Graph Convolutional Recurrent Network for Traffic Forecasting". In: (2020).

[17] Yann N Dauphin et al. "Language modeling with gated convolutional networks". In: *International conference on machine learning*. PMLR. 2017, pp. 933–941.

[18] Zili Geng et al. "STGAFormer: Spatial–temporal Gated Attention Transformer based Graph Neural Network for traffic flow forecasting". In: *Information Fusion* 105 (2024).

[19] Junqiang Wang et al. "Road Network Traffic Flow Prediction Method Based on Graph Attention Networks". In: *Journal of Circuits, Systems and Computers* 33.15 (2024).

[20] Chao 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* 34.1 (2020), pp. 914–921.

[21] Zheng Fang et al. "Spatial-temporal graph ode networks for traffic flow forecasting". In: *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 2021, pp. 364–373.