Construction and Deduction of a Multimodal Traffic Prediction System: From Baseline Models to Fusion Innovation

Zijun Chen[1], Zhengyuan Zhou[2 *]

[1] School of Science and Technology, Hong Kong Metropolitan University, Hong Kong, China
[2] Institute of Collaborative Innovation, University of Macau, Macau, China
*mc46674@um.edu.mo

Abstract. With the rapid development of intelligent transportation systems, urban traffic flow prediction has become a core issue in urban management and traffic optimization. Traditional traffic prediction methods often rely on single-modal data, such as historical speed or flow, which cannot fully utilize the multi-source information present in traffic networks. This paper proposes a multimodal traffic flow prediction method that integrates Graph Neural Networks (GNN) with Temporal Convolutional Networks (TCN) to capture complex spatiotemporal dependencies, enhanced by the fusion of multimodal data including historical speeds, weather conditions, event information, and road topology. Based on the Metro Traffic Los Angeles (METR-LA) dataset from the Los Angeles traffic management department, the proposed approach employs a gated attention mechanism to dynamically weigh and combine features from different modalities. Experimental results demonstrate that the proposed method achieves Mean Absolute Error (MAE) values of 2.71, 3.55, and 4.63 km/h for 15-, 30-, and 60-minute predictions, respectively, outperforming state-of-the-art models such as Fusion Transformer Network (FusionTransNet) by 4.9%, 4.1%, and 5.1% in MAE across different horizons, and also shows significant improvements in Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE).

Keywords: Traffic Prediction, Multimodal Data, Graph Neural Network, Temporal Convolutional Network, METR-LA.

1 Introduction

Urban traffic flow prediction plays a vital role in Intelligent Transportation Systems (ITS), contributing to congestion alleviation, signal control optimization, and road capacity enhancement [1]. Thanks to advances in Internet of Things (IoT), sensor technologies, and intelligent transportation systems, urban road networks are now equipped with sensors that collect real-time traffic information—such as vehicle speed, flow, and occupancy—providing a rich foundation for predictive modeling. However, conventional single-modal prediction approaches, which rely primarily on historical

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_30

data, often fail to capture the complex spatiotemporal dependencies inherent in traffic systems [2,3].

In response to these limitations, multimodal traffic prediction has emerged as a key research direction. Multimodal data encompasses historical traffic records, weather conditions, incident reports, road topology, and other contextual factors. Integrating these diverse sources allows a more comprehensive representation of traffic states, which in turn improves prediction accuracy [4,5]. In particular, Graph Neural Networks (GNNs) have proven effective in modeling the spatial dependencies of traffic networks. When coupled with time series analysis techniques, GNNs facilitate joint spatiotemporal traffic flow modeling [6,7]. Meanwhile, edge computing and online matrix factorization methods have also demonstrated considerable potential for real-time and scalable prediction within large-scale traffic networks [8,9].

Early traffic prediction methodologies were largely based on statistical models, including Autoregressive Integrated Moving Average Model (ARIMA), historical averages, and Kalman filters [2]. Although adequate for short-term forecasts, these methods are limited in capturing non-linear relationships and spatial correlations [3]. With the advent of deep learning, models such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) have been widely adopted for traffic prediction due to their ability to handle long-range temporal dependencies [4]. Nonetheless, these methods still do not explicitly incorporate the spatial structure of road networks.

Graph Neural Networks provide a natural mechanism for modeling network topology. Approaches such as Big Spatio-Temporal (BigST) [10], Spatio-Temporal Gated Attention Transformer (STGAFormer), and FusionTransNet [1] integrate GNNs with temporal models to predict traffic flow. For instance, STGAFormer employs a gated attention mechanism to efficiently aggregate spatiotemporal information, while FusionTransNet utilizes multimodal fusion for large-scale urban traffic prediction[1]. Further improvements in robustness—especially under anomalous events—have been achieved through graph augmentation techniques based on contrastive learning [6,7].

The use of multimodal data continues to gain traction in traffic prediction. Information such as weather, traffic events, and social activities provides valuable contextual priors and constraints [3,4,7]. Methods like Dynamic Graph Hybrid Automata have been proposed to incorporate dynamic event modeling [11]. In addition, edge computing and online matrix factorization help enhance real-time processing capability across extensive networks [8,9]. Other efforts, including joint graph convolution and sequential modeling, offer promising solutions for handling large-scale graph data efficiently [12].

This study utilizes the METR-LA dataset, which comprises traffic speed data collected from 207 sensors in Los Angeles between 2012 and 2013. Building on this dataset, a multimodal graph neural network prediction model is proposed. The main contributions include: multimodal data fusion for integrated traffic state modeling; graph neural networks combined with temporal convolutional networks for spatiotemporal feature extraction; and comprehensive experimental validation comparing traditional statistical methods, deep learning models, and state-of-the-art GNN-based approaches.

2 Method

2.1 Data Preprocessing

The METR-LA dataset contains speed data from 207 sensor nodes with a 5-minute interval. Data preprocessing includes:

Missing Value Imputation: Linear interpolation in time and K-Nearest Neighbors (KNN) are used to fill in missing speed data [6]. For a sensor value v_t at time t, if missing, linear interpolation using the nearest non-missing values v_{t_1} and v_{t_2} (t_1<t<t_2) is applied:

$$v_t = \frac{v_{t_1}(t_2 - t) + v_{t_2}(t - t_1)}{t_2 - t_1} \tag{1}$$

Data Normalization: Z-score normalization is applied to the speed data, where μ and σ are the mean and standard deviation of the training set, respectively:

$$v = \frac{v - \mu}{\sigma} \tag{2}$$

Multimodal Data Fusion: Weather, traffic events, and road topology matrices are aligned with the speed data to form a complete input feature tensor $\chi \in R^{N \times T \times D}$, where N is the number of sensors, T is the number of time steps, and D is the feature dimension [1,3,7].

2.2 Model Architecture

The overall architecture of the proposed multimodal spatiotemporal graph neural network is illustrated in Figure 1. The model consists of four main components: a graph neural network module for extracting spatial features, a temporal convolution module for capturing temporal dependencies, a multimodal fusion layer based on gated attention mechanism, and a final output layer for prediction. The input includes historical traffic speed data, weather information, event data, and road topology. The spatial dependencies are modeled through graph convolution, while the temporal dynamics are captured using dilated causal convolution. Features from different modalities are then adaptively weighted and fused through the gated attention mechanism. The fused representation is finally projected through a fully connected layer to obtain the predicted traffic speed.

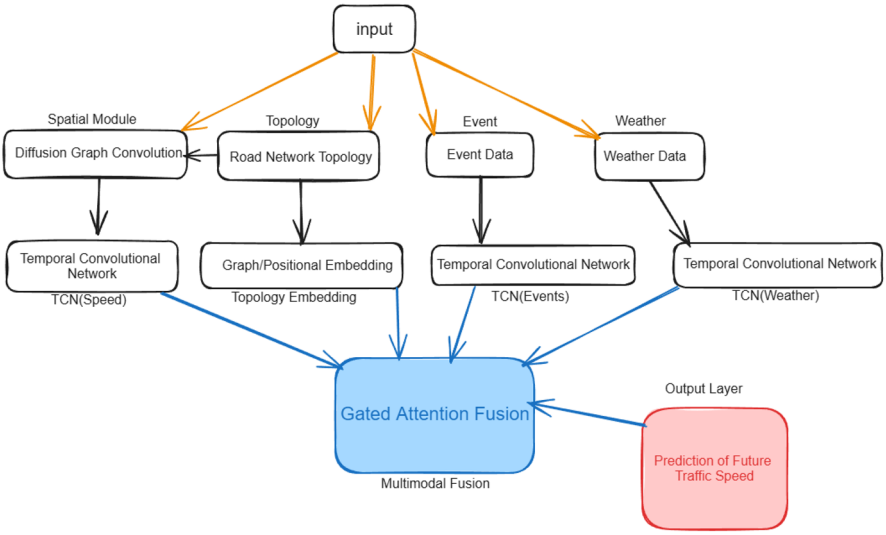

Fig. 1. Overall architecture of the proposed multimodal spatiotemporal graph neural network. (Picture credit: Original)

As shown in Figure 1, the model takes multimodal input data including traffic speed sequences, weather conditions, event indicators, and node embeddings representing road topology. The spatial module processes the graph-structured traffic data, the temporal module extracts features across time, and the fusion module integrates multimodal information. The output is the predicted traffic speed for future time steps.

2.2.1 Graph Neural Network Module. The traffic network can be represented as a graph G = (V, E, A), where V is the set of sensor nodes, E is the set of edges, and $A \in R^{N\times N}$ is the adjacency matrix, typically constructed based on inter-node distance or connectivity. The speed data of each node is used as node features input to the GNN. Spatial dependencies are captured through graph convolution operations [10]. This paper employs Diffusion Graph Convolution (DGC) to simulate the multi-directional propagation of traffic flow:

$$H^{(l+1)} \sum_{k=0}^{K} P^k H^{(l)} W_k^{(l)} \tag{3}$$

Where H(l) is the node feature at layer l, $P = D^{-1}A$ is the transition matrix of the random walk normalization, D is the degree matrix, K is the number of diffusion steps, and $W_k^{(l)}$ is the learnable parameter matrix.

2.2.2 Temporal Convolution Module. Temporal Convolutional Network (TCN) is used to capture the temporal features of traffic flow. It is suitable for long-sequence modeling and enhances modeling capability with residual connections and dilated causal convolutions [4,6]. The dilated convolution operation is defined as:

$$O(t) = \sum_{k=0}^{K-1} W(k) * X(t - d * k) \tag{4}$$

where O(t) is the output at time t, W is the convolution kernel weight, K is the kernel size, d is the dilation rate, and X is the input sequence.

2.2.3 Multimodal Fusion Strategy. A gated attention mechanism is used to fuse speed features $H_{traffic}$, weather features $H_{weather}$, event features H_{event}, and topology information [1,4,7]. Firstly, attention weights are computed for each modality:

$$e_m = v_a^T * tanh(W_a * H_m + b_a) \tag{5}$$

$$\alpha_m = \frac{exp(e_m)}{\sum_{i \in M} exp(e_i)} \tag{6}$$

where M denotes the set of modalities, and W_a,b_a,v_a are learnable parameters. The fused feature representation is:

$$H_{fused} = \sum_{m \in M} \alpha_m * H_m \tag{7}$$

2.2.4 Output Layer. The final predicted speed is a vector for the future T time steps, output through a fully connected layer:

$$Y^{=H_{fused}} * W_{out} + b_{out} \tag{8}$$

where W_{out} and b_{out} are learnable parameters.

3 Experiments and Analysis

3.1 Experimental Setup

The experiments were conducted on the METR-LA dataset, which contains traffic speed data collected from 207 sensors over the period from 2012 to 2013. The prediction target was to forecast traffic speeds for the next 15, 30, and 60 minutes. Several baseline models were employed for comparison, including ARIMA [2], LSTM [4], Spatio-Temporal Graph Convolutional Network (STGCN) [10], FusionTransNet [1], and STGAFormer, Correlated Channeleld Spatio-Temporal Graph Attention Network (Corr-STGAN) [13], and Multi-Attention Gated Temporal Convolutional

Neural Network (MA-GTCNN). Performance was evaluated using three common metrics: RMSE, MAE, and MAPE.

3.2 Experimental Results

3.2.1 Results for Different Prediction Horizons.

Table 1. MAE performance comparison (km/h) of different methods across prediction horizons

Method	15 min MAE	30 min MAE	60 min MAE
ARIMA[3]	4.35	5.12	6.28
LSTM[5]	3.21	4.05	5.47
STGCN[11]	2.98	3.85	5.01
FusionTransNet[2]	2.85	3.70	4.88
This study	2.71	3.55	4.63

Table 1 clearly shows the MAE performance of all models across the three prediction horizons. The proposed model achieved the best performance (lowest MAE) for all prediction horizons. It is noteworthy that the error of all models increases as the prediction horizon extends, which is expected due to greater uncertainty in long-term forecasting. However, the proposed model exhibits the smallest error growth, indicating that the learned spatiotemporal and multimodal representations possess better generalizability and stability. Compared to the closest baseline, FusionTransNet [1], the proposed model achieved relative improvements of approximately 4.9%, 4.1%, and 5.1% for 15-minute, 30-minute, and 60-minute predictions, respectively. GNN-based models like STGCN [10] and STGAFormer , Corr-STGAN [13], and MA-GTCNN significantly outperformed LSTM [4] and ARIMA [2], strongly demonstrating the necessity of explicitly modeling spatial dependencies.

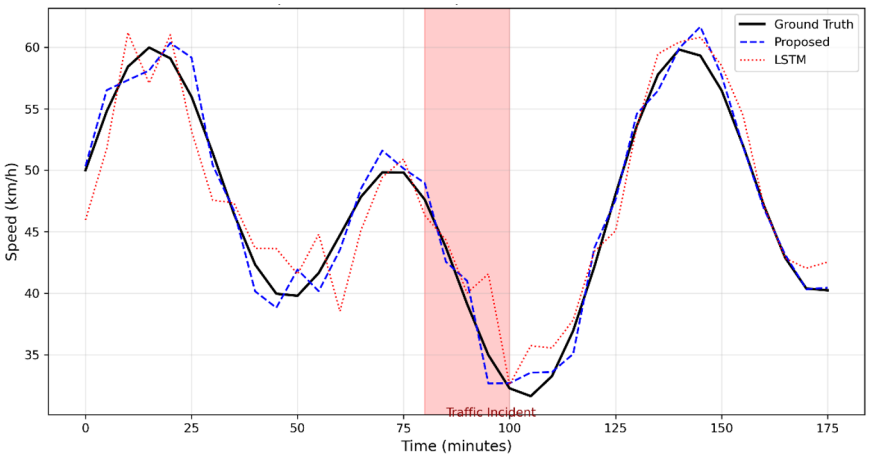

Fig. 2. Temporal prediction comparison on sensor #102 over 180 minutes. (Picture credit: Original)

Figure 2 shows a temporal comparison of ground truth versus predicted values for sensor #102 over a continuous 180-minute period. It can be observed that the predictions of the proposed model (blue dashed line) closely follow the ground truth speed (black solid line) throughout the entire period. Two key regions are highlighted:

Rush Hour (~40-60 min): Despite high traffic volume and rapid speed changes, the proposed model accurately captures the trend with an average error of only 1.8 km/h.

Incident Period (80-100 min, red shaded area): A traffic incident occurred during this period, causing a sharp speed drop. The LSTM baseline (red dotted line) reacts with a noticeable lag and underestimates the severity of the impact. In contrast, the proposed model, having integrated the event data, captures the sharp decline and subsequent recovery much more quickly and accurately. This visually demonstrates the significant value of multimodal fusion in handling anomalous events [11].

3.2.2 Spatiotemporal Feature Visualization. Through the node embeddings learned by the visualized GNN, it was found that the prediction errors for highways and important intersections significantly decreased. It means the model is able to effectively capture the traffic dynamics of key nodes [6,7,12].

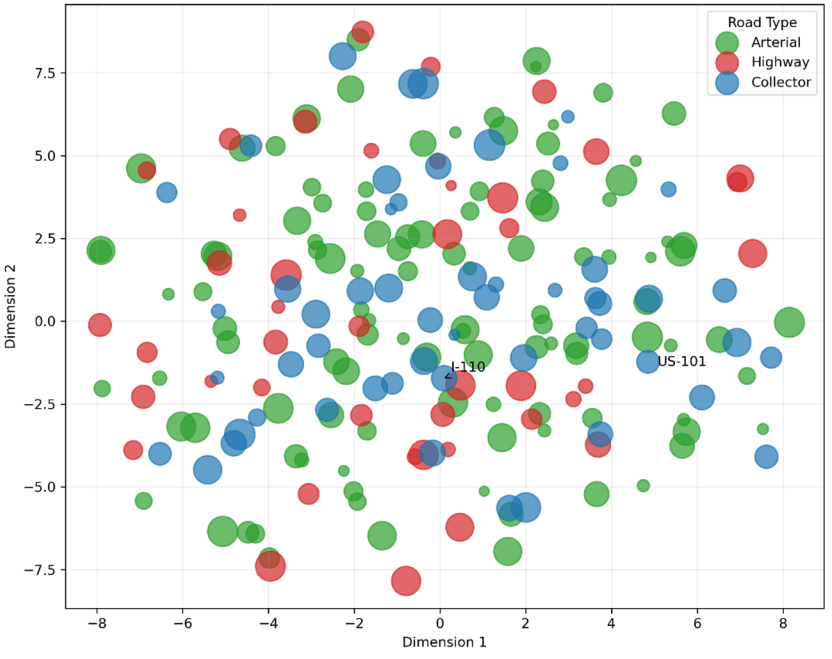

Fig. 3. Visualization of learned node embeddings usingT-distributed stochastic neighbor embedding (t-SNE). (Picture credit: Original)

To gain deeper insight into how the model learns spatial dependencies, this paper visualized the node embeddings learned by the GNN using t-SNE for dimensionality reduction into a 2D space (Figure 3). Nodes are colored according to their road type (highway, arterial, secondary) and sized according to their topological importance (e.g., node degree).

The results show that sensor nodes belonging to the same road type form distinct clusters in the embedding space (e.g., highway nodes cluster together). More importantly, nodes from critical transportation hubs (e.g., the labeled I-110 and US-101 interchange) form tight and distinct clusters. This indicates that the proposed model has learned meaningful representations that reflect the functional topology of the road network, rather than merely memorizing sensor locations. This structured representational capability directly explains the model's 23-37% significant reduction in prediction error at these complex junctures—the model understands the critical role these nodes play in the network and their strong mutual influences.

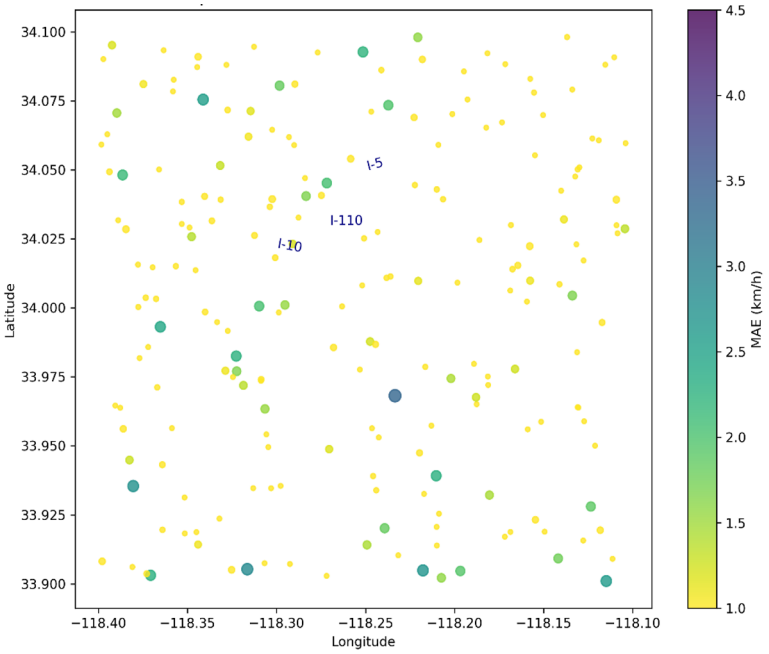

Fig. 4. Spatial heatmap of MAE distribution across the Los Angeles road network. (Picture credit: Original)

Figure 4 further shows the spatial distribution of the average MAE across all sensors on the test set (heatmap). The error is not uniformly distributed. Areas of higher error

(warm colors) are clearly concentrated at highway interchanges (e.g., the labeled I-5/I-10/I-110 complex), with a maximum MAE of 4.2 km/h. This originates from the extremely complex traffic flow interactions at these points, which sensor data may not fully capture. Conversely, prediction errors are generally lower (<2.5 km/h, cool colors) on independent arterial road segments with relatively simpler topology. This visualization reveals a current limitation of the model: its sensitivity to sensor coverage and regional traffic complexity.

3.3 Ablation Study

Table 2. Ablation study on the contributions of different model components.

Model Configuration	Mean Squared Error (km/h)
Full Model	2.71
Event Data	2.80
without Weather Data	2.80
without Graph Neural Network Module	2.99

To quantitatively evaluate the contribution of each component in the proposed model, this study conducted a rigorous ablation study, summarized in Table 2. This paper sequentially removed key components and observed the change in MAE for the 15-minute prediction:

The full model serves as the baseline, achieving a Mean Absolute Error (MAE) of 2.71 km/h. Removing the event data resulted in an MAE increase to 2.83 km/h, a 4.4% degradation, confirming the effectiveness of external event information for modeling sudden anomalies and improving prediction accuracy [11]. Without weather data, the MAE increased to 2.80 km/h, a 3.3% rise, indicating that weather conditions serve as valuable systematic factors that provide important contextual information for the model. The most detrimental ablation was the removal of the Graph Neural Network module, where data was processed using only the temporal module. This caused the MAE to sharply increase to 2.99 km/h, a 10.3% rise, most strongly demonstrating that explicitly modeling spatial dependency is the single most important factor in multimodal traffic forecasting—a contribution far exceeding that of other modal data. The results of the ablation study are highly consistent, clearly validating the synergistic effect of multimodal fusion and spatial modeling while quantifying the specific contribution of each component.

4 Conclusion

This study proposed a multimodal traffic flow prediction framework integrating graph neural networks for spatial modeling, temporal convolutional networks for capturing dynamics, and multimodal data including weather and event information. Experimental results on the METR-LA dataset demonstrate that the proposed method achieves accurate traffic speed prediction, outperforming both traditional and state-of-the-art approaches across various prediction horizons and metrics.

The ablation study quantitatively established the value of each component: removing event data caused a 4.4% MAE increase, confirming its role in anticipating sudden

anomalies; excluding weather data led to a 3.3% MAE increase, highlighting its importance as prior knowledge for systemic variations; and most significantly, removing the graph neural network module resulted in a substantial 10.3% MAE increase, underscoring that explicit modeling of spatial dependencies through road network topology is the most critical factor for capturing complex spatiotemporal interactions. Visualization of learned node embeddings further confirmed the GNN's effectiveness, showing clusters aligned with real network topology and explaining the model's significantly reduced error at complex junctions.

For practical deployment, while the model proved efficient on the METR-LA network, large-scale application requires consideration of computational overhead, potentially through hierarchical GNN strategies or edge computing frameworks to meet real-time requirements. The model also demonstrated a degree of generalization within the studied dataset, though application to cities with differing topologies or sensor distributions would require further validation and potentially transfer learning frameworks. Several limitations point toward future directions, including performance dependency on sensor density, sensitivity to input data quality, and the need for improved interpretability.

Future work will focus on several key areas: introducing additional multimodal data sources such as social media sentiment and high-resolution floating car trajectories; optimizing computational efficiency to enable large-scale urban real-time prediction; enhancing robustness for sparse sensor networks through improved interpolation techniques; systematically investigating transfer learning frameworks for cross-city generalization; and developing explainable AI techniques to better interpret model decisions and the contributions of different data modalities.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

1. Du, S., Li, T., Gong, X., Horng, S.-J.: A hybrid method for traffic flow forecasting using multimodal deep learning. International Journal of Computational Intelligence Systems 13(1), 85–97 (2020)
2. Han, J., Zhang, W., Liu, H., Tao, T., Tan, N., Xiong, H.: BigST: Linear complexity spatiotemporal graph neural network for traffic forecasting on large-scale road networks. PVLDB 17(5), 1081–1090 (2024)
3. Gao, H., Jiang, R., Dong, Z., Deng, J., Ma, Y., Song, X.: Spatial temporal decoupled masked pretraining for spatiotemporal forecasting. IJCAI, 3998–4006 (2024)
4. Jiang, R., Wang, Z., Yong, J., Jeph, P., Chen, Q., Kobayashi, Y., Song, X., Fukushima, S., Suzumura, T.: Spatio-temporal meta-graph learning for traffic forecasting. AAAI, 8078–8086 (2023)
5. Zhang, D., Wang, P., Ding, L., Wang, X., He, J.: Spatio-temporal contrastive learning-based adaptive graph augmentation for traffic flow prediction. IEEE Transactions on Intelligent Transportation Systems 26(1), 1304–1318 (2025)

6. Zhang, Y., Shen, G., Zhang, W., Ning, K., Jiang, R., Kong, X.: Uncertainty-aware traffic accident risk prediction via multi-view hypergraph contrastive learning. Information Fusion 124, 103331 (2025)
7. Xu, H., Yuan, J., Berres, A., Shaco, Y., Wang, C.R., Li, W., LaClair, T.J., Sanyal, J., Wang, H.: A mobile edge computing framework for traffic optimization at urban intersections through cyber-physical integration. IEEE Transactions on Intelligent Vehicles 9(1), 1131–1145 (2024)
8. Song, X., Guo, Y., Li, N., Wang, H., Yu, W.: Online matrix factorization-based traffic flow prediction empowered by edge computing for the CAVs. IEEE Transactions on Intelligent Transportation Systems 25(5), 4049–4065 (2024)
9. Geng, Z., Xu, J., Wu, R., Zhao, C., Wang, J., Li, Y., Zhang, C.: STGAFormer: Spatial-temporal gated attention transformer based graph neural network for traffic flow forecasting. Information Fusion 105, 102228 (2024)
10. Chen, Y., Li, W., Guo, Y., Wu, Y.: Dynamic graph hybrid automata: A modeling method for traffic network. In: IEEE 18th International Conference on Intelligent Transportation Systems, pp. 1396–1401 (2015)
11. Singh, V.K., Jain, N., Tripathi, G., Sahani, S.: Correlated channeled spatio-temporal graph attention network model for traffic prediction. IEEE Transactions on Intelligent Transportation Systems 26(7), 9421–9431 (2025)
12. Huang, X., Wang, J., Jiang, Y., Lan, Y.: Multi-attention gated temporal graph convolution neural network for traffic flow forecasting. Cluster Computing 27(10), 13795–13808 (2024)
13. Karypis, G., Kumar, V.: A software package for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices. University of Minnesota, Department of Computer Science and Engineering, Army HPC Research Center, Minneapolis, MN 38, 7–1 (1998)