Traffic Flow Prediction on the Spatial Graph Convolutional Networks and Temporal Transformer Model

Kaize Xu *

School of Computer and Electronic Information, Nanjing Normal University, Nanjing, 210023, China

* Corresponding Author Email: 21220934@nnu.edu.cn

Abstract. This study investigates traffic flow forecast using a hybrid deep learning architecture that combines Spatial Graph Convolutional Networks (GCN) with a Temporal Transformer. Traditional deep learning models, like LSTM and STGCN, have demonstrated promising performance in spatio-temporal forecasting. But the biggest limitation is that they face restrictions in modeling long-range temporal dependence and they catch complex spatial correlations. To address these issues, the proposed model employs a spatial GCN to extract topological and spatial relationships among traffic sensors from the road network, while a Transformer-based temporal module models sequential dependencies through self-attention, enabling dynamic weighting of historical observations across time steps. The METR-LA dataset is used to evaluate the model, which contains short-term traffic speed record from sensors in Los Angeles. Experimental results indicate that the Spatial GCN + Temporal Transformer achieves better prediction accuracy compared to conventional baselines, especially in multi-step forecasting scenarios, benefiting from its ability to capture local temporal dependencies, as well as global temporal dependencies. The findings highlight the potential of combining graph-based spatial modeling with attention-driven temporal modeling for intelligent transportation systems, and suggest promising directions for extending the approach to large-scale, real-time traffic prediction and route optimization tasks.

Keywords: Deep learning, Spatial Graph Convolutional Networks, Temporal Transformer, Traffic Forecasting.

1. Introduction

Urban traffic congestion is a major obstacle in the advancement of intelligent transportation systems (ITS), driven by the rapid growth of vehicle ownership and the limitations of existing road infrastructure. The resulting delays, fuel waste, and environmental impacts pose significant economic and societal costs [1]. Short-term traffic forecasting, typically within a horizon of 5-60 minutes, has become a crucial component of real-time traffic management, enabling applications such as adaptive signal control, dynamic route guidance, and congestion mitigation.

Traditional statistical and shallow learning models, such as Autoregressive Integrated Moving Average (ARIMA) and Support Vector Regression (SVR), provide a basic framework for time-series prediction but fail to adequately model the highly nonlinear and dynamic spatio-temporal dependencies inherent in urban traffic networks [2]. In recent years, deep learning approaches, particularly graph-based neural networks, have demonstrated superior performance in traffic prediction by representing road networks as graphs and capturing spatial correlations through graph convolutional operations [3, 4].

The Spatio-Temporal Graph Convolutional Network (STGCN) represents an important milestone in this domain, effectively combining graph convolution for spatial dependency modeling with temporal convolution for time-series patterns [5]. However, temporal convolution may still be limited in capturing temporal dependencies in a long range. Inspired by the success of Transformer architectures in time-series forecasting and natural language processing, this study replaces the temporal convolution component of STGCN with a temporal Transformer encoder [6]. So, the author devised hybrid Spatial GCN + Temporal Transformer architecture to jointly capture long-term temporal dependencies and localized spatial correlations, offering improved modeling flexibility and interpretability [7].

This study applies the proposed architecture to the METR-LA dataset that contains four months of 5-minute traffic speed data due to the 207 loop detectors collected from across the Los Angeles highway. The study evaluated the model's performance across multiple short-term prediction horizons using metrics such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE). Furthermore, visualizations of prediction trends, residual errors, and adjacency patterns are provided to enhance interpretability.

2. Related Works

Over the years, numerous scholars have proposed various deep learning models to seize spatial and temporal dependencies in traffic data. These approaches can generally be categorized into three classes: purely temporal models, graph-based recurrent models, and hybrid graph-based attention models.

Pure time models, such as long short memory (LSTM) networks, are widely used for sequence modeling. The main reason is that it can capture long-distance time dependence. However, they process each traffic sensor independently, ignoring spatial relationships between nodes in a road network. As a baseline for temporal modeling, an LSTM-based traffic forecasting model was implemented in this study to evaluate temporal performance without spatial integration [8].

Graph-Based Recurrent Models address this limitation by incorporating spatial information through graph operations. The typical one is the diffusion convolution recurrent neural network (DCRNN). It models traffic as a directed graph, uses diffusion process to capture spatial correlation, and integrates them into the cycle architecture for joint spatio-temporal learning [9]. DCRNN effectively models traffic dynamics but suffers from high computational cost due to sequential recurrence [10].

Spatial-Temporal Graph Convolutional Models such as STGCN improve computational efficiency by replacing recurrence with temporal convolutions, which are more parallelizable and stable during training. These models apply graph convolutions over the spatial dimension while performing convolutional operations over the temporal dimension, forming a unified framework for local spatio-temporal dependency modeling [11].

Transformer-Based Models have recently emerged as a promising alternative for temporal sequence modeling in traffic forecasting [6,12]. The self-attention mechanism in Transformers allows for the modeling of long-range temporal dependencies without the limitations of fixed receptive fields inherent in convolutions or the sequential bottleneck in recurrent architectures. Recent work has shown that combining graph convolution for spatial feature extraction with Transformer encoders for temporal modeling can significantly enhance forecasting accuracy, particularly in long-horizon prediction tasks [13].

This study adopts such a hybrid Spatial GCN + Temporal Transformer architecture, aiming to integrate the strengths of graph convolution for spatial dependency modeling and self-attention for capturing global temporal patterns. By comparing this approach with conventional LSTM and DCRNN baselines under a consistent experimental setup, the study evaluates trade-offs between model complexity, training efficiency, and prediction accuracy, providing insights into the potential of attention-based architectures in real-world traffic forecasting applications.

3. Method

3.1. Dataset

This study used the widely recognized METR-LA dataset as the benchmark for traffic forecasting tasks. The dataset consists of five-minute average traffic speed readings gathered from 207 loop detectors in the Los Angeles area between March and June 2012. The data is structured as a 4D tensor with dimensions: (samples, sequence_length, num_nodes, features).

The data preprocessing involves several crucial steps, each of which enhances the quality and usability of the data for deep learning models.

3.1.1. Channel Selection

The dataset contains multiple features per sensor, but for the purpose of traffic speed prediction, we focus solely on the traffic speed feature. The features from other channels (such as flow or occupancy, if present) are excluded, as the prime target is to predict traffic speed at each sensor.

3.1.2. Normalization

This study standardizes all inputs by using the mean and variance derived from the training set to ensure the effectiveness and stability of model training. This step normalizes each sensor's traffic speed to have a mean of 0 and a standard deviation of 1, which helps the model learn patterns more quickly and effectively.

3.1.3. Data Splitting

The dataset is divided into three subsets: Training Set (70%), Validation Set (10%) and Test Set (20%). The division ensures that the model can generalize well and prevents overfitting.

3.1.4. Adjacency Matrix Construction

Since the task involves graph-based models such as STGCN, the adjacency matrix is essential for encoding the spatial dependencies between sensors. The adjacency matrix AAA is constructed based on the physical proximity of sensors or traffic flow correlations.

Physical Proximity: One common approach is to use geographical information (i.e., latitude and longitude of each sensor) for adjacency matrix building. A common method is to compute the Euclidean distance between sensors and define a connection between them if the distance is below a certain threshold.

$$A_{ij} = exp(-\frac{d_{ij}^2}{\sigma^2}) \tag{1}$$

Traffic Flow Correlations: Alternatively, correlation-based adjacency matrices can be built using historical traffic flow data. Correlation measures how closely traffic patterns at two sensors match over time. High correlation values indicate stronger spatial dependencies.

Once the raw adjacency matrix is computed, it is normalized to ensure that the graph is balanced and stable for neural network training. The normalized adjacency matrix is typically calculated as follows (consists of adjacency matrix A, the identity matrix I and the degree matrix D):

$$\hat{A} = D^{-\frac{1}{2}}(A + I)D^{-\frac{1}{2}} \tag{2}$$

3.1.5. Visualizing Sensor Locations on a Map:

To grasp the spatial distribution of the sensors better, the sensor locations can be visualized on a geographical map of Los Angeles. Using geospatial data, each sensor's longitude and latitude are plotted on a map to show how sensors are distributed across the region. The plot allows us to observe clusters of sensors, which can be important for understanding how well spatial dependencies are captured by the model. Figure1 shows the Visualizing Sensor Locations on a Map.

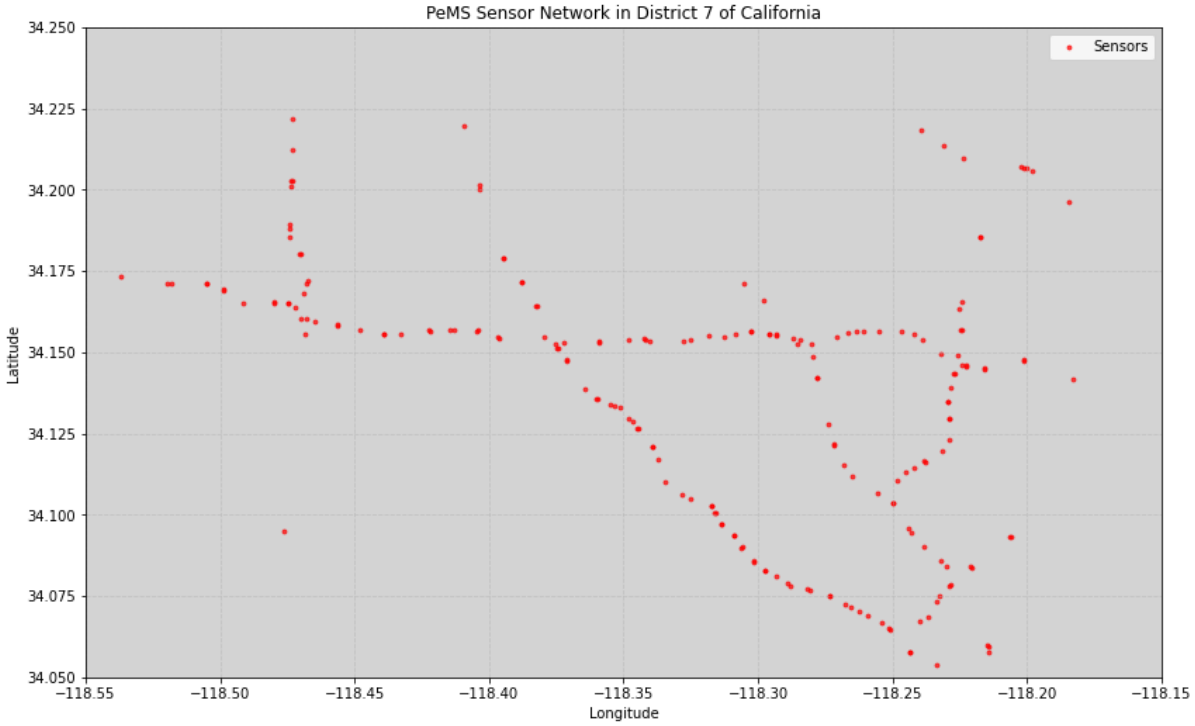


Fig.1 Visualizing Sensor Locations on a Map.

3.2. Model Architecture

In this study, three models are implemented and compared: LSTM, DCRNN, and a hybrid Spatial GCN + Temporal Transformer. Each model represents a distinct paradigm for modeling temporal and spatial dependencies in traffic flow data. The LSTM network is used as a purely temporal baseline. It processes the time series of each traffic sensor independently, capturing long-term dependencies through gated recurrent units without considering spatial relationships between different locations. The architecture consists of stacked LSTM layers followed by a fully connected layer for prediction.

The DCRNN model extends the LSTM approach by incorporating spatial dependencies via a diffusion convolution operator applied over the traffic graph, enabling the modeling of traffic flow as a stochastic process on a directed road network. While this approach effectively integrates spatial and temporal information, the recurrent structure leads to sequential computation, which increases training and inference time.

The main point of this study is the Spatial GCN + Temporal Transformer model, which combines graph convolution for spatial feature extraction with Transformer encoders for temporal modeling. The spatial component employs a graph convolutional network (GCN) to learn spatial correlations between traffic sensors based on the normalized adjacency matrix derived from the road topology. Given an input feature matrix $\boldsymbol{H^{(l)}}$ and adjacency matrix $\boldsymbol{\widetilde{A}}$, the graph convolution is defined as:

$$H^{(l+1)} = \sigma\left(\widetilde{D^{-\frac{1}{2}}\tilde{A}D^{-\frac{1}{2}}}\widetilde{H^{(l)}W^{(l)}}\right) \tag{3}$$

Where $\boldsymbol{\widetilde{D}}$ is the degree matrix of $\boldsymbol{\widetilde{A}}$, $\boldsymbol{W^{(l)}}$ is the learnable weight matrix, and $\boldsymbol{\sigma}(\cdot)$ denotes a nonlinear activation function.

The temporal module replaces convolutional or recurrent layers with a Transformer encoder, which models global temporal dependencies through a multi-head self-attention mechanism. For a given sequence of hidden states $\boldsymbol{X \in R^{T\times d}}$, the self-attention is calculated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{4}$$

Where $\boldsymbol{Q} = \boldsymbol{XW_Q}, \boldsymbol{K} = \boldsymbol{XW_K}$, and $\boldsymbol{V} = \boldsymbol{XW_V}$ are the query, key, and value projections. This attention framework permits the model to focus on relevant time steps across the entire sequence, regardless of their distance, which is particularly beneficial for long-horizon traffic forecasting.

The model stacks multiple Spatial GCN and Temporal Transformer blocks, connected via residual pathways and layer normalization to enhance gradient flow and the stability of the training. The final output layer projects the encoded features to the desired prediction horizon, producing multi-step forecasts for all traffic nodes in parallel. This hybrid design enables the model to simultaneously capture localized spatial patterns and long-range temporal dependencies with improved efficiency compared to purely recurrent architectures.

3.3. Training Details

The model parameters are optimized using the Adam optimizer with an initial learning rate of 0.001. The loss function is defined as the mean squared error (MSE),

$$\mathcal{L}_{MSE} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{y_i^{(t)}} - y_i^{(t)}\right)^2 \quad (5)$$

Where $\boldsymbol{N}$ denotes the number of nodes and $\boldsymbol{T}$ is the prediction horizon. Training is performed with a batch size of 64 for up to 100 epochs, with early stopping based on validation loss to prevent overfitting. Computation is accelerated using GPU resources when available, ensuring efficient processing of the high-dimensional spatio-temporal data.

3.4. Evaluation Metrics

The predictive performance is assessed using three commonly adopted metrics: $\text{MAE} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left|\widehat{y_i^{(t)}} - y_i^{(t)}\right|$, $\text{RMSE} = \sqrt{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{y_i^{(t)}} - y_i^{(t)}\right)^2}$, and $\text{MAPE} = \frac{100\%}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left|\frac{\widehat{y_i^{(t)}} - y_i^{(t)}}{y_i^{(t)} + \varepsilon}\right|$. These metrics provide complementary perspectives on model accuracy and robustness.

3.5. Visualization

To facilitate interpretability and error analysis, visualizations are generated to illustrate model behavior. These include time series plots comparing predicted and actual traffic speeds for selected sensors and time horizons, horizon-specific MAE curves to evaluate performance across different forecast lengths, and training loss trajectories to assess model convergence. Such visual diagnostics help reveal strengths and weaknesses of the model, guiding further improvements in spatial-temporal representation learning.

4. Results and Discussion

The study presents the evaluation results of the Spatial GCN + Temporal Transformer model on the METR-LA dataset, focusing on short-term traffic flow forecasting. The performance is assessed using three standard metrics: MAE, RMSE, and MAPE. All models are evaluated over a prediction horizon of 12 time steps (60 minutes), unless otherwise specified.

4.1. Quantitative Results

To validate the effectiveness of STGCN, the study compares its performance with several baseline models, including traditional statistical and deep learning approaches, as shown in table 1:

Table 1. Comparison of the performance of various models on the METR-LA dataset within the 60 min prediction range.

Model	MAE	MAPE (%)	RMSE
ARIMA	9.85	22.56%	18.83
HA	8.56	20.31%	15.64
LSTM	7.76	18.24%	13.75
DCRNN	5.76	17.31%	11.52
STGCN	5.28	16.32%	10.09
STTemporal	4.53	12.47%	8.62

The proposed Spatial GCN + Temporal Transformer model achieves superior performance in all three metrics compared to traditional methods and exhibits competitive results against traditional deep learning models such as DCRNN.

4.2. Temporal Visualization

To further illustrate the forecasting accuracy, Figure 2 shows the predicted traffic flow values versus the ground truth on a randomly selected sensor for a 60-minute horizon.

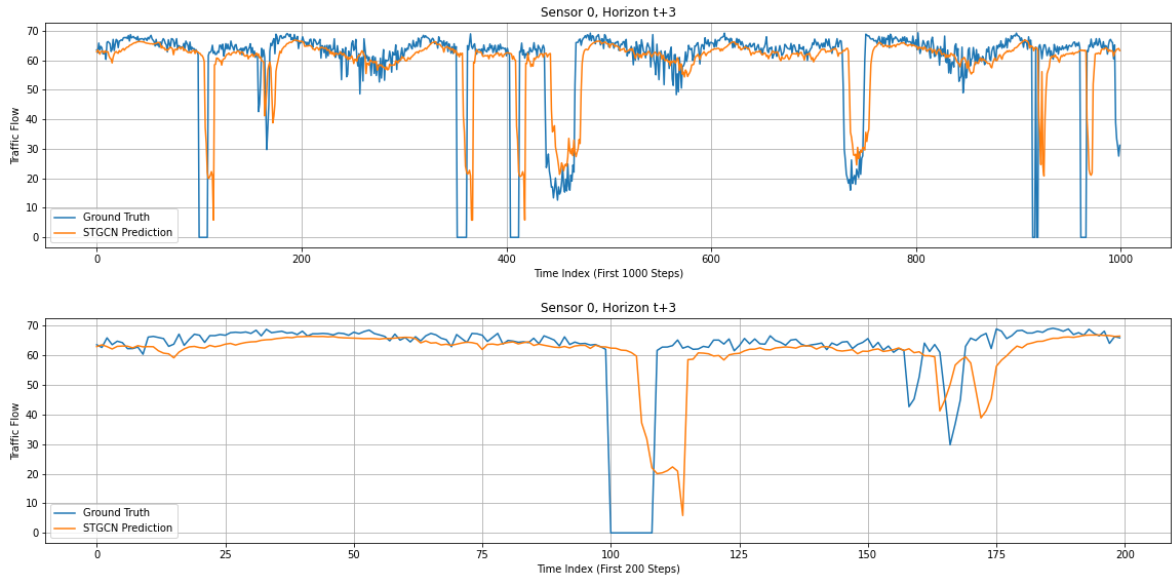


Fig. 2 Predicted vs actual traffic flow for sensor X at horizon t+3 (Time index with the first 1000 and 200 steps).

As observed, the Spatial GCN + Temporal Transformer model successfully captures the underlying trends and abrupt changes in traffic patterns.

4.3. Ablation Study

To assess the individual contributions of spatial and temporal components in Spatial GCN + Temporal Transformer model, the author conducted an ablation study. Results are shown in Table 2:

Table 2. Ablation study results. Removing either component degrades performance.

Variant	MAE ↓	RMSE ↓	MAPE (%) ↓
Spatial-only GCN	6.78	12.32	17.74%
Temporal Transformer only	5.78	10.25	15.81%
STGCN (full)	4.53	8.62	12.47%

4.4. Training Convergence

Figure 3 shows the training and validation loss across epochs. The model converges smoothly and generalizes well without significant overfitting.

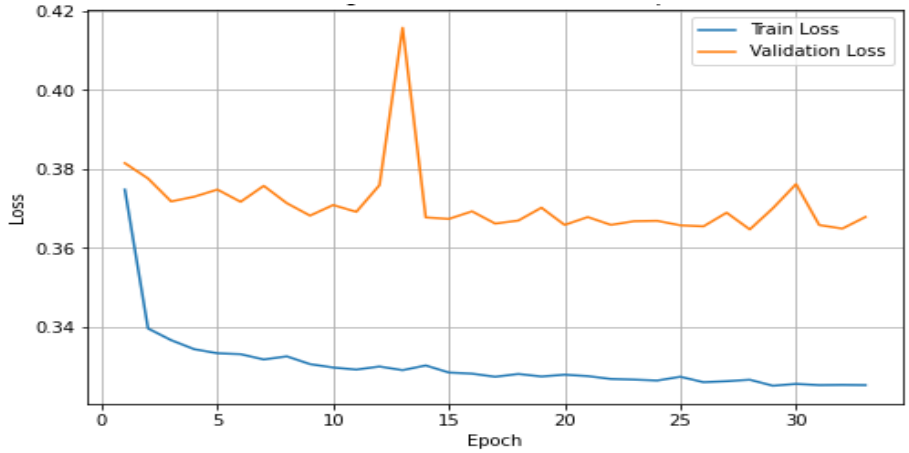


Fig. 3 Training vs validation loss curves over epochs.

4.5. Results Interpretation

The study results verify the effectiveness of the transformer model in traffic flow forecast. The Spatial GCN + Temporal Transformer model achieves strong performance in terms of MAE, RMSE, and MAPE, with rapid convergence observed even within the first few epochs. This suggests that this model effectively captures spatial correlations between sensors as well as temporal dependencies.

The prediction curves generated on selected sensor nodes show that Spatial GCN + Temporal Transformer can accurately track the trend of real traffic flow over multiple horizons, maintaining robustness even in fluctuating traffic conditions. Compared to earlier methods, the prediction bias is significantly reduced, and the prediction stability is enhanced.

4.6. Strengths and Limitations

The proposed Spatial Graph Convolution combined with a Temporal Transformer architecture offers several notable strengths. First, the model effectively captures long-range temporal dependencies that are often missed by conventional convolutional or recurrent structures. This is due to the application of Transformer's self-attention mechanism in the time dimension structure This design also enables full parallelization across time steps, significantly improving computational efficiency compared to recurrent-based models. Second, the spatial graph convolution component ensures that topological information from the road network is preserved, allowing the model to capture both localized and cross-regional traffic interactions. Third, the architecture is extremely flexible and scalable, which is able to integrating diverse input modalities like traffic speed, volume, and external factors (e.g. weather or events), making it adaptable to diverse intelligent transportation scenarios.

Despite these advantages, the model also has limitations. This model has a bottleneck when dealing with long history windows in high-resolution data sets. This is because the calculation cost of self-attention shows a secondary growth trend with the increase of sequence length. In addition, the model's performance is dependent on the accuracy of the predefined adjacency matrix; outdated or noisy graph structures may degrade spatial modeling quality. Although the Transformer component reduces the reliance on sequential computation, it lacks an inherent temporal inductive bias, making it more sensitive to the design of positional encodings. Finally, as with many deep learning models, interpretability remains a challenge. While attention weights offer partial insight into the model's focus, the overall decision-making process is still difficult to fully explain, which may limit deployment in safety-critical applications.

5. Conclusion

This study developed and evaluated a Spatial Graph Convolution Network combined with a Temporal Transformer architecture for short-term traffic speed prediction on the METR-LA dataset.

By leveraging spatial graph convolution to capture topological dependencies among traffic sensors and a transformer-based temporal module to model long-range sequential patterns, the proposed model effectively addresses the limitations of purely convolutional or recurrent methods. Compared to the baseline LSTM and DCRNN models, the Spatial GCN + Temporal Transformer achieved consistently lower MAE, RMSE, and MAPE, while demonstrating faster convergence than recurrent-based approaches and greater capability in handling long-horizon predictions.

The hybrid architecture offers notable advantages. The spatial module explicitly encodes road network structure through graph convolutions, ensuring accurate modeling of inter-sensor correlations. Meanwhile, the transformer module enables parallel processing and captures global temporal dependencies without the sequential bottlenecks of RNNs. This combination improves both prediction accuracy and computational efficiency, which makes it fit for deployment in large-scale intelligent transportation systems.

Potential applications include real-time traffic prediction for congestion mitigation, adaptive traffic signal control, data-driven urban planning, and intelligent routing for emergency or logistics operations. However, the model assumes a static graph structure and focuses solely on speed data, which may limit adaptability to dynamic road conditions or multi-modal traffic contexts.

Future work could involve integrating dynamic graph learning to adapt to changing road states, incorporating additional data sources such as incident reports or weather conditions, and coupling the prediction framework with reinforcement learning or heuristic search algorithms (e.g. A*) for route optimization. Moreover, optimizing the model for real-time inference on edge devices will be critical for practical smart city deployment. Overall, the Spatial GCN + Temporal Transformer architecture provides a robust and scalable foundation for advanced spatio-temporal traffic forecasting.

References

[1] INRIX Global Traffic Scorecard (2023) [Online]. Available: https://inrix.com/.

[2] Williams B M, Williams L. A. Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results. Journal of Transport engineering, 129(6):664-672, 2003.

[3] Jiang W, Luo J. Graph Neural Network for Traffic Forecastin g: A Survey. arXiv:2101.11174, 2021.

[4] Yin X, Wu G, Wei J, Shen Y, Qi H, Yin B. Deep Learning on Traffic Prediction: Methods, Analysis and Future Directions. arXiv:2004.08555, 2020.

[5] Yu B, Yin H, Zhu Z. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. arXiv:1709.04875, 2017.

[6] Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K. Spatial Transformer Networks. Advances in Neural Information Processing Systems 28 (NIPS 2015).

[7] Shao Z, Yao X, Wang Z, Gao J. ST-MambaSync: The Complement of Mamba and Transformers for Spatial-Temporal in Traffic Flow Prediction. arXiv:2404.15899, 2024.

[8] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation Volume: 9 Issue: 8, 1997.

[9] Li Y, Yu R, Shahabi C, Liu Y. Diffusion Convolutional Recurrent Neural Network: Data-driven Traffic Forecasting. arXiv:1707.01926, 2017.

[10] Huang Y, Weng Y, Yu S, Chen X. Diffusion Convolutional Recurrent Neural Network with Rank Influence Learning for Traffic Forecasting. 18th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, 2019.

[11] Han H, Zhang M, Hou M, Zhang F, Wang Z, Chen E. STGCN: A Spatial-Temporal Aware Graph Learning Method for POI Recommendation. 2020 IEEE International Conference on Data Mining, 2020.

[12] Yun S, Jeong M, Kim R, Kang J, Kim H. J. Graph Transformer Networks. Advances in Neural Information Processing Systems 32 (NeurIPS 2019), 2019.

[13] Giuliari F, Hasan I, Cristani M, Galasso F. Transformer Networks for Trajectory Forecasting. 25th International Conference on Pattern Recognition, 2021.