scientific reports

Check for updates

OPEN

Multi-scale spatial-temporal transformer for traffic flow prediction

Zicheng Qiu[1], Han Wu[1]✉, Guoqing Teng[1], Hao Wu[2], Zhengyu Huang[1] & Meng Zhao[1]

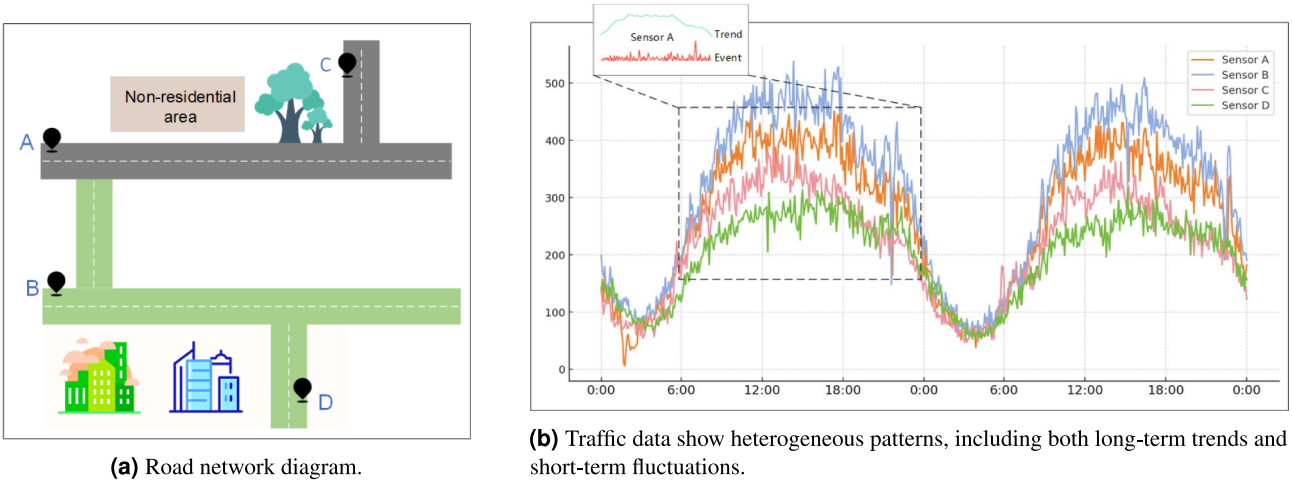
Traffic flow prediction is a critical component of Intelligent Transportation Systems (ITS), playing a vital role in improving travel efficiency and enhancing road safety. However, achieving accurate prediction remains challenging due to complex spatial-temporal dependencies, particularly spatial heterogeneity and multi-scale temporal patterns. To address these challenges, we propose a novel framework called Multi-Scale Spatial-Temporal Transformer (MSSTFormer). It integrates a two-stage spatial attention module to extract spatial features, which handles the complex relationships between global and local dependencies and enhances interactions among strongly correlated key nodes. In this way, the model's ability to capture critical spatial relationships is significantly improved. Additionally, the frequency dual-channel attention module in the model, a novel module, can enhance the model's ability to capture complex temporal dynamics by decomposing multi-scale temporal features and separately modeling long-term trends and short-term fluctuations through the interplay between high-frequency and low-frequency components. Furthermore, the model incorporates a gated mechanism in the data embedding layer to eliminate redundant information, thereby optimizing input data quality. Experimental results on four public traffic datasets demonstrate that MSSTFormer outperforms existing models on most datasets, showing improvements in prediction accuracy. Moreover, we enhance model interpretability by visualizing the learned frequency dual-channel attention weights. Our code is available at https://github.com/whaaaa123/MSSTFormer.

Keywords Traffic flow prediction, Attention mechanism, Spatial–temporal, Frequency domain

With the rapid acceleration of urbanization, intelligent transportation systems (ITS) play a crucial role in alleviating traffic congestion, optimizing resource allocation, and enhancing travel efficiency. As one of the core technologies of ITS, traffic flow prediction leverages historical traffic data to accurately forecast current traffic volumes, thereby providing essential support for traffic planning and management[1]. Currently, the primary challenge in traffic flow prediction lies in effectively extracting spatial-temporal features, in order to improve prediction accuracy and better support traffic management and planning.

Early studies predominantly employed time series analysis models, such as the Autoregressive Integrated Moving Average (ARIMA)[2] and Autoregressive (VAR) models[3], to forecast traffic flow. However, these models perform poorly in complex traffic flow forecasting due to their inability to capture nonlinear characteristics and leverage the spatial properties of urban networks. With advances in deep learning, Graph Neural Networks (GNN) and self-attention mechanisms have gradually emerged as mainstream methods for traffic flow prediction[4–6]. GNN-based methods typically focus on local structures to capture dependencies between neighboring nodes. Nevertheless, in urban traffic systems, local modeling often fails to fully reflect the dynamics of overall traffic flow[7]. In contrast, the self-attention mechanism supports global modeling through fully connected interactions that dynamically adjust inter-node relationship weights to capture dependencies across the entire network. However, due to the spatial heterogeneity of transportation networks, this fully connected approach may fail to capture regional differences effectively. For example, nearby node pairs may exhibit different traffic trends due to variations in functional areas, while nodes farther apart but sharing similar functions may exhibit similar traffic patterns. As illustrated in Fig. 1a, sensors A, B, and C are deployed along the road, with sensors A and B positioned adjacently, while sensor C is situated at a non-adjacent location. From a spatial correlation perspective, sensors A and B would be expected to exhibit similar traffic flow trends, whereas A and C might differ. However, due to urban zoning regulations, sensors A and C, both located in non-residential areas, exhibit more similar traffic fluctuations, whereas sensor B, situated in a residential zone, displays markedly different traffic patterns from A.

[1]School of Computer Science and Engineering, Chongqing University of Science and Technology, Chongqing 401331, China. [2]School of Electrical and Electronic Engineering, North China Electric Power University, Beijing 100096, China. ✉email: wuhan@cqust.edu.cn



(a) Road network diagram.

(b) Traffic data show heterogeneous patterns, including both long-term trends and short-term fluctuations.

Fig. 1. The findings about traffic data.

From a temporal perspective, traffic flow data typically exhibits pronounced multi-scale characteristics. As shown in Fig. 1b, the traffic flow sequence from sensor D involves multiple complex temporal patterns, including prominent long-term trends and significant short-term fluctuations[5]. Long-term trends generally reflect daily, weekly, and seasonal periodicities in travel behaviors, whereas short-term fluctuations result mainly from unexpected events or sudden congestion. However, existing traffic flow forecasting methods usually focus merely on single-scale temporal features[8–12]. Although stacked convolutional networks[13,14] can capture both long-term and short-term temporal dependencies to some extent, they rely on a single temporal scale and fail to differentiate between multi-scale temporal characteristics. This limitation restricts the model's ability to effectively learn multi-scale temporal features.

To address the above issues, this paper proposes a Multi-Scale Spatial-Temporal Transformer (MSSTFormer) for traffic flow prediction. In the data embedding layer, we introduce a gating mechanism to filter redundant information and enhance the quality of the input data. Moreover, we design a two-stage spatial attention module, which integrates global spatial modeling with a key-node enhancement strategy. Additionally, a frequency dual-channel module approach is proposed, which decouples the low-frequency and high-frequency temporal features, independently modeling long-term trends and short-term fluctuations. This enhances the model's capability to capture complex temporal dependencies. The key contributions of this study are summarized as follows:

- This paper proposes the MSSTFormer model, which integrates a spatial-temporal self-attention mechanism for traffic prediction.It effectively addresses the challenges of spatial heterogeneity and multi-scale temporal dynamics.
- We introduce a gating mechanism into the data embedding stage to suppress redundant information and improve input quality.
- A two-stage spatial attention module is designed to capture global spatial dependencies during the spatial feature extraction phase and reinforce interactions among strongly correlated nodes via a key mask matrix, thereby capturing the complex relationships between global and local spatial features.
- A frequency dual-channel temporal attention module is proposed, decoupling traffic flow time series into low-frequency and high-frequency components to enhance the model's ability to learn temporal features, enabling more effective differentiation between long-term trends and short-term fluctuations.
- MSSTFormer has been extensively evaluated on four real-world road traffic datasets, demonstrating superior performance compared to most state-of-the-art baseline models.

Related works

Traditional traffic forecasting

Traditional traffic flow prediction methods can be categorized into statistical methods, machine learning methods, and deep learning methods. These approaches analyze traffic flow variations to support traffic management and planning.

Statistical methods were dominant in early traffic flow prediction research. They treated traffic flow as a linear problem and used fixed theoretical models for forecasting. Regression functions were predefined, and parameters were determined by processing raw data. Predictions were made using these regression functions, such as time series models[15] and Kalman filter models[16]. Although these models are computationally simple, they cannot handle the nonlinear dynamics of traffic flow and lack robustness when faced with sudden events. With the increasing complexity and nonlinearity of traffic flow data, researchers have gradually shifted their focus toward machine learning techniques. Machine learning methods overcome the limitations of statistical models as they do not require strict assumptions about data distribution. These approaches utilize data-driven techniques to uncover underlying patterns. Typical methods include Bayesian networks, Support Vector

Machines (SVM)[17], and k-nearest neighbors (KNN)[18]. These methods overcome several limitations of statistical models and perform better in complex traffic environments. However, they still require substantial manual effort for feature extraction.

With the rapid development of deep learning, many new methods have started to explore how to extract richer features from time-series data and effectively model spatial-temporal dependencies. However, despite significant progress in spatial-temporal modeling, most methods still rely on traditional time-domain modeling (such as LSTM and GRU). While these methods can capture long-term dependencies in time series, they typically learn direct relationships between time steps, lacking explicit modeling of frequency components. For instance, Huang et al. proposed a method combining deep belief networks to automatically extract high-level features from traffic data[19], thus eliminating the need on manually defined features inherent in traditional methods. Lv et al. applied sparse autoencoders to traffic flow prediction[20], which automatically extract key features while maintaining model sparsity. With the advancement of computing hardware, models like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) have been widely adopted for modeling temporal dependencies[21]. However, LSTM and GRU models cannot handle predictions with unequal input and output lengths. Ilya Sutskever et al. introduced the Seq2Seq model[22]. Convolutional Neural Networks (CNNs) capture dependencies in Euclidean space using convolution kernels. The Temporal Convolutional Network (TCN) enhances time feature extraction capabilities[23]. As research progresses, multi-model fusion has become more prominent. Wu et al. proposed a deep learning framework that combines CNN and LSTM[24]. Yu et al. introduced the spatial-temporal Recursive Convolutional Network (SRCN), which integrates the advantages of deep convolutional neural networks (DCNNs) and LSTM networks. This network is capable of simultaneously predicting both long-term and short-term traffic conditions[25]. Despite their effectiveness in capturing spatial-temporal characteristics, most methods are designed for Euclidean data and face challenges in adapting to complex topological relationships in traffic networks. At the same time, researchers have gradually recognized that time series data inherently possess multi-scale features, and learning multi-scale temporal features can significantly enhance the model's ability to capture the dynamics of time series. For example, Chen et al. proposed a multi-scale convolutional network (MS-CNN) to capture temporal information at different scales, which was successfully applied to time series classification tasks[26]. Additionally, Chen et al. also proposed a time-aware multi-scale recurrent neural network (MS-RNN), which can adaptively select the most important temporal scale features[27]. Zhong et al. introduced the multi-scale decomposition MLP-Mixer, which explicitly decomposes the input time series into different components for analysis. Although these methods effectively capture multi-scale features, they are typically limited to time-domain modeling and rely on predefined scales, which makes them somewhat limited when handling dynamically changing multi-scale time series data.

Graph convolutional networks

Graph Convolutional Networks effectively model the spatial structure of traffic network data by sharing information between nodes via graph convolution operations, improving traffic flow prediction accuracy. Zhao et al. proposed Temporal Graph Convolutional Networks (T-GCN)[9], combining GCN with GRU to capture spatial and temporal dependencies. Bai et al. introduced an Adaptive Graph Convolutional Recurrent Network (AGCRN)[28]. This model learns node-specific patterns and infers dependencies in traffic sequences automatically. Li et al. proposed trajectory-based Graph Neural Networks (TrGNN)[29], which integrate vehicle trajectory data and environmental information into the road graph network. Although these studies recognize the importance of capturing both spatial and temporal features, most focus on spatial dependencies and overlook dynamic dependencies that evolve over time.

In recent years, with increasing focus on dynamic spatial-temporal features, several new methods have emerged. Wu et al. introduced Graph WaveNet[30], which dynamically learns the adjacency matrix to capture real spatial dependencies without relying on a fixed graph structure. Peng et al. proposed Dynamic Graph Convolutional Networks (DGCN)[31], which dynamically update the graph structure. This captures spatial dependencies at different time points and under varying traffic conditions. Huang et al. presented Time-Varying Graph Convolutional Networks (TVGCN)[32], which capture spatial-temporal features at multiple scales and dynamically update the graph structure. With the rise of multi-view learning, more researchers have applied this approach to traffic flow prediction. For example, Bai et al. proposed Multi-Task Synchronous Graph Neural Networks (MTSGNN)[33], which learn dynamic spatial dependencies across different tasks simultaneously. This enhances the prediction of regions transitions. Wang et al. introduced Multi-View Bidirectional Spatial-Temporal Graph Neural Networks (BiSTGN)[34], which construct three spatial-temporal graph sequences from different time-related perspectives, enhancing the model's ability to capture multi-dimensional temporal information. Huang et al. further developed Multi-View Dynamic Graph Convolutional Networks (MV-DGCN)[35], combining multiple views and dynamic graph convolutional networks (DGCN) to capture spatial-temporal features and adapt to dynamic changes in traffic flow. This significantly improves the representation of spatial-temporal features and traffic flow prediction accuracy.

Attention mechanism

The attention mechanism has gained widespread application in recent studies. It has an inherent global receptive field and can capture dynamic dependencies in traffic data effectively. Initially proposed by Vaswani et al.[36], the attention mechanism has been widely applied in natural language processing, machine vision, time series forecasting, and other tasks. For instance, Cheng et al. proposed a Graph Multi-Head Attention Network (GMAN)[6], which combines graph neural networks with a multi-head attention mechanism. This model enhances traffic flow prediction accuracy by jointly modeling spatial and temporal dependencies with a multi-scale attention mechanism. Huang et al. introduced a Multi-Relation Synchronous Graph Attention Network (MS-GAT)[37], which learns spatial, temporal, and channel interactions in a unified and synchronized manner for

traffic coupling. Feng et al. developed an Adaptive Graph Spatial-Temporal Transformer Network (ASTTN)[38], which jointly models spatial-temporal correlations and local spatial-temporal attention. This network employs an adaptive graph structure and multi-layer spatial-temporal attention stacking to capture spatial-temporal dependencies. Jiang et al. proposed PDFormer[39], introducing geographical and semantic spatial masks into the attention mechanism to capture both short-range and long-range dynamic spatial dependencies. Additionally, it employs a delay-perception feature transformation module to account for propagation delays in real traffic roads. Recently, Li et al. developed DDGFormer[40], a model combining a self-attention module with distance and direction awareness. It also employs a dynamic adaptive graph convolution module, enabling more effective capture of dynamic spatial-temporal dependencies.

However, to the best of our knowledge, the aforementioned models do not capture multi-scale temporal dynamics from a frequency-domain perspective. Moreover, these methods have limited ability to capture spatial heterogeneity. To this end, this paper proposes the MSSTFormer model, which separates high and low frequencies from a frequency-domain perspective to capture multi-scale temporal dynamics. Additionally, a two-stage spatial attention mechanism is introduced to better capture dynamic spatial-temporal correlations.

Problem definition

Traffic flow prediction aims to forecast the traffic volume of a transportation system at a future time point, based on historical data. In this study, we represent the road network as a graph $G = (V, E, A)$, where $V = \{v_1, \ldots, v_N\}$ is a set of N nodes, $E \subseteq V \times V$ represents the set of edges, and A is the adjacency matrix of the network. In traffic flow modeling, nodes generally represent sensors or monitoring points along the roads. Edges represent the topological connections between roads. Based on this graph, we describe the dynamic traffic flow across the entire road network using a traffic flow tensor. Here $X_t \in \mathbb{R}^{N \times C}$ represents the traffic flow state of all N nodes at time t. C denotes the feature dimension of the traffic flow. Since this study focuses solely on traffic flow prediction, the feature dimension $C = 1$.

Given the observed traffic flow tensor X of the transportation system, the goal of traffic flow prediction is to learn a mapping function f. This function maps traffic flow observations from the past T time steps to the future T' time steps:

$$[X(t-T+1), \ldots, X_t; G] \xrightarrow{f} [X(t+1), \ldots, X(t+T')] \tag{1}$$

Methodology

Figure 2 illustrates the framework of MSSTFormer, comprising a data embedding layer, stacked L spatial-temporal encoder layers, and an output layer. Each module is described in detail in the following sections.

Data embedding layer

In traffic flow prediction tasks, the data embedding layer integrates multi-dimensional features to generate a comprehensive representation. To optimize feature fusion, a gating mechanism is introduced to suppress redundant information and highlight key features, thereby improving the quality of the input data. Initially, the raw input X is transformed into a high-dimensional feature $E_f \in \mathbb{R}^{T \times N \times d}$ via a fully connected layer, where d represents the embedding dimension.

In the temporal dimension, traffic flow exhibits distinct daily and weekly periodic patterns. To capture these patterns, we model the daily and weekly time segments separately: daily periodicity $d(t)$ divides the day into 1440 minute-long time slots, and weekly periodicity $w(t)$ divides the week into 7 days. Additionally, we incorporate traffic flow data from the previous day, which is subsequently transformed via a linear layer to generate the embedding representation of these periodic features, denoted as $E_t \in \mathbb{R}^{T \times N \times d}$.

Simultaneously, to capture spatial correlations among sensors in the road network, structural information is incorporated. The graph Laplacian eigenvector method is employed to generate a spatial embedding matrix $E_s \in \mathbb{R}^{N \times d}$. This matrix maps graph node relationships into Euclidean space, thereby capturing spatial dependencies among sensors and preserving the global topology of the road network. Furthermore, a time position encoding $E_p \in \mathbb{R}^{T \times d}$ is introduced to maintain the invariance of the temporal positional information. By stacking these embeddings, X_E is obtained:

$$X_E = E_f + E_t + E_s + E_p \tag{2}$$

A gating mechanism is introduced with the aim of mitigating redundancy arising from feature fusion. The computation is defined as follows:

$$X' = W_c \left(W_a(X_E) \odot \mathrm{silu}(W_b(X_E)) \right) \tag{3}$$

where W_a , W_b , and $W_c \in \mathbb{R}^{d \times d}$. X' is subsequently fed into the spatial-temporal encoder. For convenience, we use X to represent X' in the following sections.

Spatial-temporal encoder layer

The spatial–temporal encoder layer consists primarily of two components, employing a two-stage spatial attention module for spatial feature extraction and a frequency dual-channel attention module for temporal feature processing, all within a multi-head attention framework.

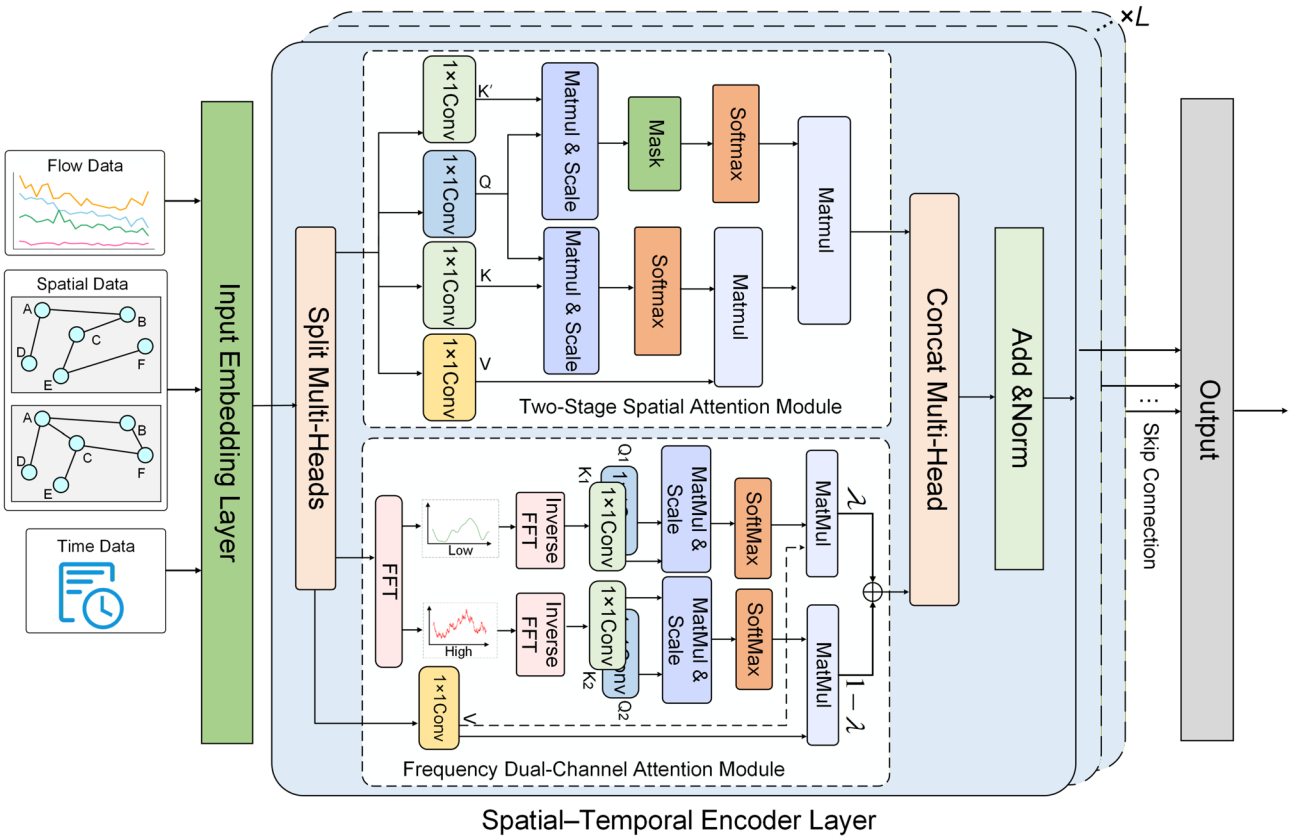


Fig. 2. The framework of MSSTFormer.

Two-stage spatial attention module

Traditional spatial self-attention mechanisms assume that each node interacts equally with all other nodes, treating the spatial graph as fully connected. However, in real-world traffic networks, spatial relationships between nodes are heterogeneous. In certain cases, traffic flow patterns between adjacent nodes may differ significantly, while traffic states at remote nodes may exhibit greater similarity. Therefore, focusing exclusively on geographically adjacent nodes might overlook potential remote spatial dependencies. To tackle this challenge, a two-stage spatial attention module is proposed. In the first stage, a global self-attention mechanism is employed to model the entire traffic network, capturing spatial-temporal dependencies across the entire network. In the second stage, a key mask matrix is constructed to reinforce the interactions between node pairs with strong correlations, while further revealing the complex dependency structure between global and local spatial domains.

In the first stage, a global self-attention mechanism is employed to model spatial dependencies within the traffic network, thereby revealing the latent dynamic spatial structures among nodes.The query matrix Q, key matrix K, and value matrix V are calculated to provide a global perspective. At time t, the traffic flow observations are multiplied by the corresponding weight matrices to obtain the query matrix Q, key matrix K, and value matrix V:

$$Q_t^{(S)} = X_{t::}W_Q^S \quad K_t^{(S)} = X_{t::}W_K^S \quad K'^{(S)}_t = X_{t::}W_{K'}^S \quad V_t^{(S)} = X_{t::}W_V^S \tag{4}$$

where, W_Q^S, W_K^S, $W_{K'}^S$, and $W_V^S \in \mathbb{R}^{d\times d'}$ denote learnable parameters, whered' represents the dimensionality of the query, key and value matrices.

The dot product between the query and key matrices is computed, followed by normalization. The spatial dependencies, denoted as $A_t^{(S)}$ (i.e., the attention scores) among all nodes at time t are computed:

$$A_t^{(S)} = \frac{Q_t^{(S)}(K_t^{(S)})^\top}{\sqrt{d'}} \tag{5}$$

The computed attention scores are then multiplied by the value matrix to produce the output of the global spatial self-attention module $G_t^{(S)}$:

$$G_t^{(S)} = \mathrm{softmax}(A_t^{(S)})V_t^{(S)} \tag{6}$$

In the second stage, owing to spatial heterogeneity, only a subset of node pairs in traffic networks exhibit significant interactions. Consequently, we propose a key mask matrix that selectively reinforces interactions between geographically proximate and semantically similar nodes.

For each node pair (i, j), we first evaluate their spatial proximity using the road network topology. If the physical distance between them is less than a predefined threshold λ , they are considered spatially connected, and the corresponding mask entry is assigned $M_{ij} = 1$. Yet spatial dependencies in traffic systems are not solely governed by physical distance. Some distant nodes may still exhibit semantic similarity due to analogous traffic dynamics. For such pairs, we measure traffic similarity using Dynamic Time Warping (DTW)[41]. We select the K nodes with the highest similarity scores $\mathrm{Sim}(i, j)$ for each node as its semantic neighbors, and assign $M_{ij} = 1$ for these selected pairs. Otherwise, $M_{ij} = 0$.

With the constructed key mask matrix $M \in \mathbb{R}^{N\times N}$, the spatial attention scores $A_t'^{(S)}$ are computed as:

$$A_t'^{(S)} = \frac{Q_t^{(S)}(K_t'^{(S)})^\top}{\sqrt{d'}} \odot M \tag{7}$$

where, $\odot$ denotes element-wise multiplication.

Finally, our approach effectively captures the intricate spatial dependencies by merging global information with local details, yielding the output of the two-stage spatial attention module, denoted as $Z_s \in \mathbb{R}^{T\times N\times d_s}$, formulated as:

$$Z_s = \mathrm{softmax}(A_t'^{(S)})G_t^{(S)} \tag{8}$$

Frequency dual-channel attention module

Traffic flow data exhibit complex multi-scale temporal dynamics, encompassing long-term trends (e.g., daily, weekly, and seasonal periodicities) and short-term fluctuations (e.g., unexpected events or abrupt congestion). However, most existing time series modeling methods rely on single-scale temporal features, neglecting the interactions between different frequency domains[8,10,11].

To better capture such interdependencies, this paper proposes a frequency dual-channel attention module that utilizes the Fourier transform to decompose time series into low-frequency components representing stable periodic trends and high-frequency components corresponding to sudden events. The method adaptively models long-term trends and short-term fluctuations by decoupling the time series data.

For node *n*, a discrete Fourier transform converts the time-series components into frequency representation $X_n^{(T)}(f)$ as follows:

$$X_n^{(T)}(f) = \sum_{t=0}^{T-1} x_n(t) \cdot e^{-i2\pi ft/T}, f \in \{0, 1, ..., T-1\} \tag{9}$$

where $x_n(t)$ denotes the traffic flow observation at node *n* at time *t*, $X_n^{(T)}(f)$ represents Fourier-transformed signal in complex form, and *T* is the length of the time window.

A low-pass filter is applied to extract low-frequency components, preserving periodic elements and long-term trends while removing high-frequency noise:

$$X_n^{(T,l)}(f) = \begin{cases} X_n^{(T)}(f), & |f| \le f_c \\ 0, & |f| > f_c \end{cases} \tag{10}$$

where f_c is the cutoff frequency of the low-pass filter, controlling the range of retained low-frequency signals. High-frequency components are obtained by subtracting the low-frequency components from the original signal:

$$X_n^{(T,h)}(f) = X_n^{(T)}(f) - X_n^{(T,l)}(f) \tag{11}$$

Prior to applying the attention mechanism, the query (*Q*) and key (*K*) matrices are computed for both low-frequency and high-frequency features, thereby facilitating effective modeling in the subsequent self-attention mechanism. The low-frequency features for node *n* are calculated as follows:

$$Q_n^{(T,l)} = X_{:n:}^{(T,l)}W_Q^l \quad K_n^{(T,l)} = X_{:n:}^{(T,l)}W_K^l \tag{12}$$

Similarly, for the high-frequency components:

$$Q_n^{(T,h)} = X_{:n:}^{(T,h)}W_Q^h \quad K_n^{(T,h)} = X_{:n:}^{(T,h)}W_K^h \tag{13}$$

where $W_Q^l, W_K^l, W_Q^h, W_K^h \in \mathbb{R}^{d\times d'}$ denote learnable parameter matrices, and $X_n^{(T,l)}$ and $X_n^{(T,h)}$ represent the low-frequency and high-frequency features, respectively.

Following frequency decomposition, a dual-channel self-attention mechanism is applied to model low-frequency and high-frequency features separately. Here, *V* denotes the value matrix extracted from the time-

series data. In the low-frequency path, long-term trends are emphasized by computing the attention matrix to extract low-frequency features:

$$A_n^{(T,l)} = \text{Softmax}\left(\frac{Q_n^{(T,l)} K_n^{(T,l)\top}}{\sqrt{d'}}\right) V_n^{(T)} \tag{14}$$

In the high-frequency path, short-term fluctuations in the traffic flow are captured:

$$A_n^{(T,h)} = \text{Softmax}\left(\frac{Q_n^{(T,h)} K_n^{(T,h)\top}}{\sqrt{d'}}\right) V_n^{(T)} \tag{15}$$

Furthermore, a learnable adaptive weighting parameter is introduced into the proposed module to dynamically adjust the contributions of low-frequency and high-frequency features to traffic flow prediction, thereby improving the adaptability of the model across multiple temporal scales. The final output of the frequency dual-channel attention module, denoted as $Z_t \in \mathbb{R}^{T\times N\times d_t}$, is formulated as follows:

$$Z_t = \lambda \cdot A_n^{(T,l)} + (1-\lambda)\cdot A_n^{(T,h)} \tag{16}$$

Output layer

Each spatial-temporal encoder module incorporates a residual connection and a 1×1 convolution to project the intermediate output X_o into a residual representation $X_{sk} \in \mathbb{R}^{T\times N\times d_{sk}}$, where d_{sk} denotes the residual feature dimension. Then, we obtain the final hidden state $X_{hid} \in \mathbb{R}^{T\times N\times d_{sk}}$ by summing the outputs of each residual connection layer. For multi-step forecasting, X_{hid} is passed through an output module composed of two successive 1×1 convolutional layers, which refine and map the hidden features to the prediction space:

$$\hat{X} = \text{Conv2}(\text{Conv1}(X_{hid})) \tag{17}$$

where $\hat{X} \in \mathbb{R}^{T'\times N\times C}$ denotes the predicted output for T' future time steps.

Experiments

Datasets

The datasets used in this study were obtained from four publicly available datasets provided by the California Department of Transportation's Highway Performance Monitoring System (PeMS): PeMS03, PeMS04, PeMS07, and PeMS08. These datasets were collected from more than 39,000 traffic detectors deployed along major highways in California's metropolitan areas. The data is recorded in real time at 30-second intervals and subsequently aggregated into 5-minute intervals, encompassing metrics such as traffic flow, average speed, and lane occupancy. Table 1 presents the details of each dataset.

Baseline models

The performance of our MSSTFormer model was compared with several baseline models, including time series forecasting models, graph neural networks (GNN) models, and attention-based models.

- HA[42]: A statistical time series analysis model that uncovers dynamic relationships.
- ARIMA[2]: A method that models traffic flow as a seasonal ARIMA process, capturing periodic fluctuations and seasonal variations by incorporating seasonal differencing and parameters.
- VAR[3]: A model that integrates Granger causality with a vector autoregression framework to analyze causal relationships between nodes and capture the temporal dynamics of network traffic.
- STGCN[5]: Utilizes graph structures to model traffic flow by employing graph convolutional networks to capture spatial dependencies and temporal dynamics.
- ASTGCN[10]: Incorporates spatial and temporal attention mechanisms to dynamically adjust the weights assigned to nodes and time steps.
- STSGCN[43]: Simultaneously models spatial-temporal dependencies by introducing a spatial-temporal resonance mechanism to elucidate the complementarity and dynamic relationships between spatial and temporal variations.
- MTGNN[44]: Captures spatial-temporal dependencies in multivariate time series through the construction of graph structures using GNNs.

Dataset	Sensors	Edges	Time range	Time steps
PeMS03	358	547	09/2018–11/2018	26,208
PeMS04	307	340	01/2018–02/2018	16,992
PeMS07	883	866	05/2017–08/2017	28,224
PeMS08	170	295	07/2016–08/2016	17,856

Table 1. The detailed information of datasets.

- STFGNN[45]: Models spatial dependencies and temporal dynamics separately, sharing information across layers via inter-layer fusion and cross-layer information propagation mechanisms.
- STGODE[46]: Utilizes Ordinary Differential Equations (ODEs) to capture spatial-temporal dynamics.
- DSTAGNN[47]: Introduces a spatial-temporal-aware mechanism to adaptively model both spatial and temporal dependencies.
- GDGCN[48]: GDGCN treats multiple historical time periods as nodes in a graph and employs a dynamic graph builder to model time-varying spatial and temporal relationships.
- STIDGCN[49]: Combines dynamic graph convolutional networks with multi-perspective modeling to capture spatial-temporal heterogeneity and reveal dynamic dependencies between nodes through an interactive learning framework.
- GMAN[6]: Dynamically adjusts the graph structure and adjacency relationships to capture nonlinear interactions between nodes, employing multi-layer attention mechanisms to model spatial-temporal features.
- ASTGNN[50]: This model combines self-attention mechanisms to capture spatial-temporal dynamics, periodicity, and spatial heterogeneity. It effectively models the spatial-temporal relationships in traffic flow through dynamic graph convolution and embedding modules.
- STID[51]: Introduces a dynamic adjacency matrix to model the time-varying nature of the graph structure, capturing temporal relationships between nodes and adaptively adjusting via a heterogeneity-handling mechanism.
- PDFormer[39]: Proposes a propagation delay-aware mechanism combined with a long-term Transformer architecture to adaptively model long-term dependencies.
- STAEformer[52]: Enhances the performance of the Vanilla Transformer model by integrating sequence encoding, spatial encoding, and adaptive spatial-temporal encoding via spatial-temporal embeddings.
- DDGFormer[40]: Combines attention mechanisms with graph convolutional networks, utilizing directional and distance information to enhance traffic flow modeling capabilities.

Hyperparameter settings

All experiments are conducted on a machine equipped with an NVIDIA GeForce RTX 4090 GPU. A 5-minute time step is employed, with the input sequence length (T) set to 12 and the output sequence length (T') also set to 12. We implement MSSTFormer using Ubuntu 18.04, PyTorch 1.10.1, and Python 3.9.7. The hidden dimension d is searched over $\{16, 32, 64, 128\}$, and we selected $d = 64$. The depth of the encoder layers L is explored within $\{2, 4, 6, 8\}$, and we selected $L = 6$ based on performance considerations. The feature dimensions for both the time and space modules, d_t and d_s, are set to 32. For the PeMS03, PeMS04, and PeMS08 datasets, a batch size of 16 is used, while for the PeMS07 dataset, the batch size is set to 6. The model is trained for 200 epochs using the Adam optimizer with a learning rate of 0.001. Evaluation metrics include mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{18}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\% \tag{19}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{20}$$

Here, $\hat{y}_i$ and y_i represent the predicted and actual traffic flow values, respectively.

Performance comparison

Table 2 reports the average multi-step prediction performance across four datasets, comparing the MSSTFormer model with baseline models. The results demonstrate that MSSTFormer consistently outperforms all baseline models on most evaluation metrics. Specifically, MSSTFormer achieves a 2.12% improvement in the MAE value on the PeMS03 dataset, with the frequency dual-channel attention module making the most significant contribution to this enhancement compared to other methods. The best-performing results are highlighted in bold and the second-best results are underlined. Based on comparative analysis, several important findings are derived: 1) Compared to traditional time series models such as HA, ARIMA, and VAR, our proposed model incorporates spatial dependencies that have often been overlooked. 2) Compared to graph neural network (GNN)-based models such as STGCN, STSGCN, and GDGCN, attention mechanism-based approaches, including PDFormer and STAEformer, exhibit superior predictive performance. However, spatial modeling frameworks utilizing self-attention, such as STAEformer, do not explicitly differentiate spatial dependencies between geographically adjacent and distant nodes. In contrast, MSSTFormer designs a two-stage spatial attention module that captures global spatial dependencies while enhancing interactions among critical nodes, effectively revealing the complex relationships between global and local spatial features. 3) Among self-attention-based temporal modules, DDGFormer is one of the most competitive baseline models. It captures temporal dependencies by combining a self-attention mechanism with positional encoding. However, due to its single-scale temporal modeling approach, DDGFormer has certain limitations in capturing multi-scale temporal features in traffic flow data. In contrast, MSSTFormer proposes a frequency dual-channel attention module that

Model	PeMS03			PeMS04			PeMS07			PeMS08		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
HA(1995)	29.80	46.83	28.11	38.76	57.72	29.59	45.15	66.97	21.75	32.55	48.49	20.77
ARIMA(2003)	35.41	47.59	33.78	33.73	48.80	24.18	38.17	59.27	19.46	31.09	44.32	22.73
VAR(2016)	23.65	38.26	24.51	24.54	38.61	17.24	50.22	75.63	32.22	19.19	29.81	13.10
STGCN(2017)	17.55	30.42	17.34	21.16	34.89	13.83	25.33	39.34	11.21	17.50	27.09	11.29
ASTGCN(2019)	17.34	29.56	17.21	22.93	35.22	16.56	24.01	37.87	10.73	18.25	28.06	11.64
STSGCN(2020)	17.48	29.21	16.78	21.19	33.65	13.90	24.26	39.03	10.21	17.13	26.80	10.96
MTGNN(2020)	15.85	26.23	15.55	19.08	31.56	12.96	20.82	34.09	9.03	15.40	24.93	10.17
STFGNN(2021)	16.77	28.34	16.30	19.83	31.88	13.02	22.07	35.80	9.21	16.64	26.22	10.60
STGODE(2021)	16.50	27.84	16.69	20.84	32.82	13.77	22.59	37.54	10.14	16.81	25.97	10.62
DSTAGNN(2022)	15.57	27.21	14.68	19.30	31.46	12.70	21.42	34.51	9.01	15.67	24.77	9.94
GDGCN(2023)	14.66	**24.30**	**13.94**	18.44	29.79	12.52	20.15	33.21	8.50	14.82	23.87	9.35
STIDGCN(2024)	14.76	24.59	15.28	18.16	29.77	12.24	19.26	32.51	8.11	13.45	23.28	8.77
GMAN(2020)	16.87	27.92	18.23	19.14	31.60	13.19	20.96	34.10	9.05	15.31	24.92	10.13
ASTGNN(2021)	14.78	25.00	16.40	18.29	29.82	12.49	19.54	32.82	8.25	14.20	23.49	9.28
STID(2022)	15.33	27.40	14.79	18.60	30.91	12.36	20.62	34.00	8.86	15.00	24.70	9.50
PDFormer(2023)	14.94	25.39	15.82	18.32	29.97	12.10	19.83	32.87	8.53	13.58	23.51	9.05
STAEformer(2023)	15.35	27.55	15.18	18.22	30.18	11.98	19.14	32.60	8.01	13.46	23.25	8.88
DDGFormer(2024)	15.01	24.95	15.89	18.04	30.06	11.76	18.99	**32.25**	7.93	13.37	**23.15**	8.83
MSSTFormer(ours)	**14.35**	25.01	14.05	**17.93**	**29.40**	**11.52**	**18.79**	32.48	**7.89**	**13.20**	23.31	**8.76**

Table 2. Comparison Table of Model Performance.

decomposes temporal features into low-frequency and high-frequency components using Fourier transform, separately modeling long-term trends and short-term fluctuations. This enables the model to effectively handle variations across different time scales, thereby enhancing its predictive capability for complex traffic flow data.

- MSSTFormer w/o gate embedding: multidimensional features are fused using weighted summation without the use of a gating mechanism.
- MSSTFormer w/o two-stage spatial attention: the two-stage spatial attention is removed, and global attention is applied to extract spatial features directly.
- MSSTFormer w/o frequency dual-channel attention: low-frequency and high-frequency features are not preserved, as the attention module from the Vanilla Transformer[36] is employed for temporal modeling.

Ablation study

The MSSTFormer model consists of three primary modules: the embedding layer, the two-stage spatial attention module, and the frequency dual-channel attention module. To validate the effectiveness of each module, we conducted an ablation study on the PeMS03, PeMS04, PeMS07, and PeMS08 datasets. The experimental results are shown in Table 3. The findings are as follows: (1) Removing the gate embedding (i.e., w/o gate embedding) leads to a significant decrease in MAE, RMSE, and MAPE across all datasets, indicating that the gate embedding effectively filters unnecessary information and plays a critical role in reducing redundancy. (2) When global spatial attention is used to replace the two-stage spatial attention (i.e., w/o two-stage spatial attention), the model's performance declines significantly, particularly on the PeMS04 and PeMS07 datasets. This demonstrates that, compared to single-view spatial modeling, the two-stage spatial attention method more effectively captures complex spatial dependencies, thereby revealing the intricate spatial dynamics between global and local contexts. (3) Removing the frequency dual-channel attention (i.e., w/o frequency dual-channel attention) results in a noticeable inability to capture short-term fluctuations and long-term trends, especially on the PeMS07 dataset. This outcome shows that learning temporal correlations only through gated recurrent units is insufficient to capture multi-scale temporal features, and the frequency dual-channel attention module plays a crucial role in handling multi-scale temporal features in traffic flow data.

We evaluate the impact of each module in the MSSTFormer model on single-step (one step every 5 min) prediction evaluation metrics on the PEMS08 dataset in Fig. 3. We can observe that: (1) Incorporating the dynamics of traffic flow with the gate embedding significantly reduces prediction errors in both short-term and long-term forecasts. (2) The two-stage spatial attention mechanism improves the model's ability to capture spatial dependencies, leading to better performance compared to w/o two-stage spatial attention in both short-term and long-term predictions. (3) The inclusion of frequency dual-channel attention allows the model to effectively separate short-term fluctuations and long-term trends, outperforming the variant w/o frequency dual-channel attention.

Parameter sensitivity analysis

To further explore the impact of various parameters, we performed sensitivity analysis for the spatial mask used in MSSTFormer. Specifically, we explored various values for each hyperparameter within predefined search

Dataset	Models	Gate embedding	Two-stage spatial attention	Frequency dual-channel attention	MAE	RMSE	MAPE (%)
PeMs03	w/o gate embedding	×	✓	✓	14.48	25.73	14.24
	w/o two-stage spatial attention	✓	×	✓	14.61	26.08	14.13
	w/o frequency dual-channel attention	✓	✓	×	14.56	25.82	14.16
	MSSTFormer	✓	✓	✓	14.35	25.01	14.05
PeMs04	w/o gate embedding	×	✓	✓	18.05	30.11	11.71
	w/o two-stage spatial attention	✓	×	✓	18.18	30.46	11.59
	w/o frequency dual-channel attention	✓	✓	×	18.11	30.21	11.64
	MSSTFormer	✓	✓	✓	17.93	29.40	11.52
PeMs07	w/o gate embedding	×	✓	✓	18.92	33.18	8.11
	w/o two-stage spatial attention	✓	×	✓	19.05	33.55	7.98
	w/o frequency dual-channel attention	✓	✓	×	18.96	33.27	8.02
	MSSTFormer	✓	✓	✓	18.79	32.48	7.89
PeMs08	w/o gate embedding	×	✓	✓	13.31	24.04	8.92
	w/o two-stage spatial attention	✓	×	✓	13.42	24.33	8.84
	w/o frequency dual-channel attention	✓	✓	×	13.38	24.15	8.89
	MSSTFormer	✓	✓	✓	13.20	23.31	8.76

Table 3. Comparison of ablation experiments.

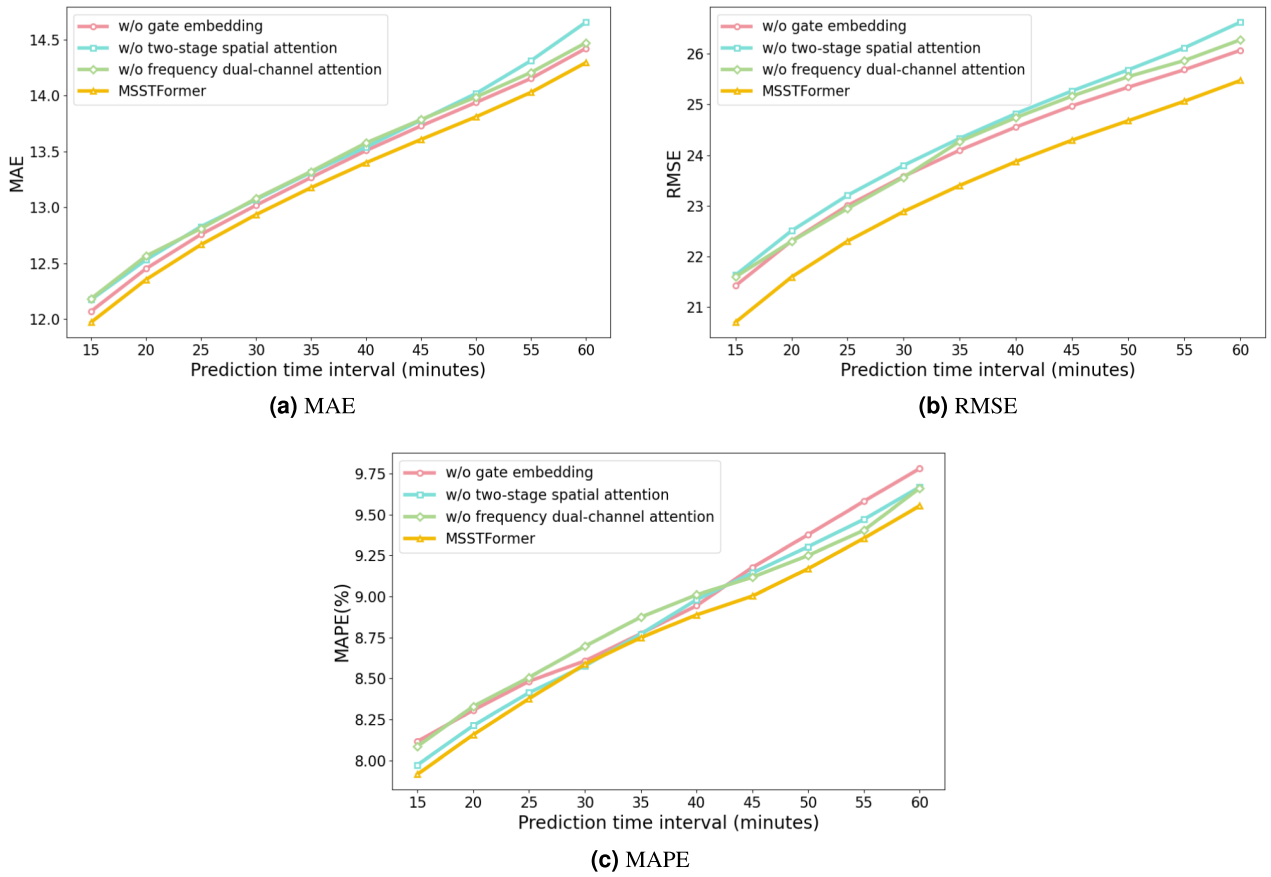


Fig. 3. Comparison of single-step prediction performance on the PeMS08 dataset.

spaces:[6–9] for the number of nearest neighbors K selected based on the DTW similarity and[4–7] for the distance threshold λ. This comprehensive analysis allowed us to evaluate the impact of different configurations on the performance of our MSSTFormer model. The results are shown in Fig. 4

From the figure, we have the following observations: (1) Increasing the number of nearest neighbors K based on DTW similarity improves model performance up to $K = 7$, beyond which further increases offer minimal

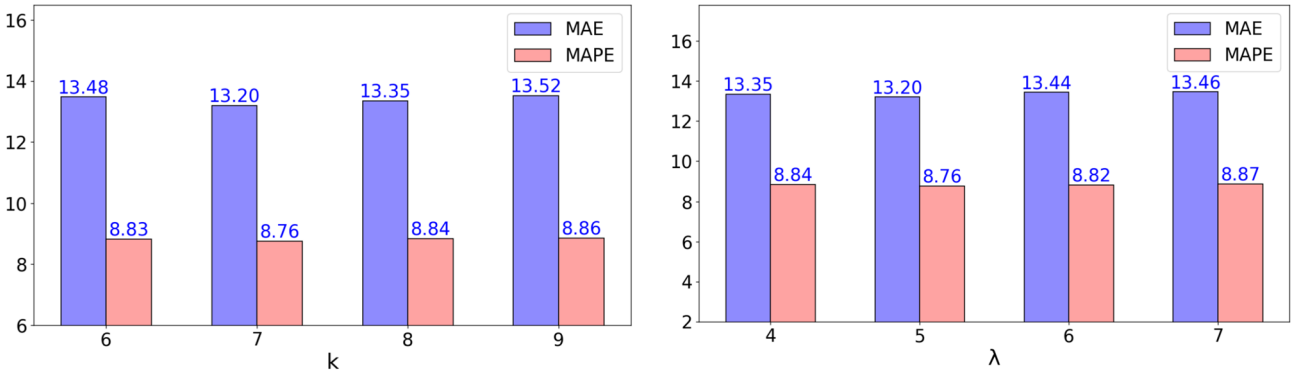
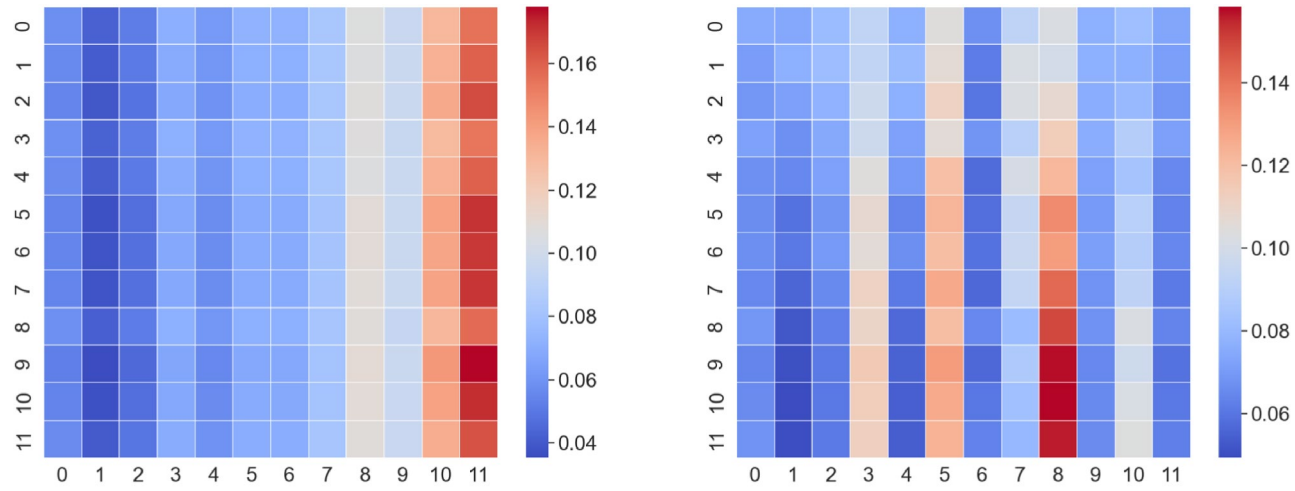


Fig. 4. Experimental results of the hyperparameter study on the PeMS08 dataset.



(a) Heatmap of attention weights in the low-frequency.

(b) Heatmap of attention weights in the high-frequency.

Fig. 5. The visual analysis of the frequency dual-channel attention module.

benefit. A value of $K = 7$ effectively captures both short-term fluctuations and long-term trends, while lower values (e.g., $K = 6$) introduce noise, and higher values (e.g., $K = 9$) reduce sensitivity to sudden changes. (2) The distance threshold $\lambda = 5$ best preserves spatial dependencies, allowing the model to capture relevant relationships without introducing unnecessary complexity. Smaller values (e.g., $\lambda = 4$) overemphasize local dependencies, while larger values (e.g., $\lambda = 7$) lead to overfitting and reduced efficiency by including irrelevant connections.

Case study

To verify the effectiveness of MSSTFormer in decomposing traffic flow data into distinct temporal patterns, we selected 12 consecutive time steps at 5-minute intervals from node 42 of the PeMS04 dataset, spanning 08:30 to 09:30 on January 8, 2018. Figure 5 illustrates the attention weight distributions across high-frequency and low-frequency channels for these selected time steps. The color gradient from blue to red in the heatmap represents the increase in attention weights, where red corresponds to higher attention and blue to lower attention. As shown in Fig. 5a, the attention weights in the low-frequency channel exhibit a smooth and progressively increasing pattern over time, which reflects steady increases in traffic flow during off-peak hours. This pattern indicates that MSSTFormer prioritizes long-term traffic flow trends rather than short-term fluctuations. In contrast, Fig. 5breveals significant fluctuations in the attention weights of the high-frequency channel, with distinct peaks observed around time steps 6–8. These peaks indicate sudden changes in traffic flow near sensor 42, likely due to congestion or incidents. This highlights MSSTFormer's ability to capture transient events, utilizing high-frequency attention to focus on short-term disruptions and improving the model's responsiveness to dynamic traffic conditions.

By separately processing these two distinct temporal features, MSSTFormer dynamically adjusts their contributions to traffic flow predictions. Specifically, during peak traffic periods, the model allocates greater attention to high-frequency features, thereby enhancing its sensitivity to short-term fluctuations. Conversely, during stable periods, it prioritizes low-frequency features ensures accurate predictions of long-term traffic

patterns. Consequently, MSSTFormer effectively distinguishes between short-term fluctuations and long-term trends, significantly improving prediction accuracy.

Conclusions

In this study, we propose a novel model named MSSTFormer. A two-stage spatial attention module is integrated, which effectively addresses the complex relationships between global and local dependencies while enhancing the interactions among strongly correlated key nodes, thereby improving the model's ability to capture intricate spatial features. Additionally, a new frequency dual-channel attention module is incorporated, which decouples high-frequency and low-frequency components to separately model long-term trends and short-term fluctuations, further enhancing the model's ability to capture complex temporal dynamics. Moreover, a gating mechanism is embedded in the data embedding layer to reduce information redundancy. Extensive experiments conducted on four real-world datasets validate the superior performance of MSSTFormer. The results demonstrate its advantage over state-of-the-art models in traffic flow prediction tasks and highlight the strong potential of MSSTFormer for real-time traffic flow prediction, particularly in environments with dynamic and unpredictable conditions. Nonetheless, the model is not yet lightweight, and deploying it in large-scale urban transportation systems may introduce considerable computational and memory overhead. In future work, we plan to enhance the model's computational efficiency and scalability to facilitate large-scale, real-time deployment. We will also focus on optimizing the model's architecture to achieve a better balance between prediction accuracy and resource consumption, ensuring its practical applicability in intelligent transportation systems.

Data availability

Our code is available at https://github.com/whaaaa123/MSSTFormer.

Received: 18 April 2025; Accepted: 19 December 2025
Published online: 25 December 2025

References

1. Tedjopurnomo, D. A., Bao, Z., Zheng, B., Choudhury, F. M. & Qin, A. K. A survey on modern deep neural network for traffic prediction: Trends, methods and challenges. *IEEE Trans. Knowl. Data Eng.* **34**, 1544–1561. https://doi.org/10.1109/TKDE.2020.3001195 (2022).
2. Isufi, E., Loukas, A., Simonetto, A. & Leus, G. Autoregressive moving average graph filtering. *IEEE Trans. Signal Process.* **65**, 274–288. https://doi.org/10.1109/TSP.2016.2614793 (2017).
3. Lu, Z., Zhou, C., Wu, J., Jiang, H. & Cui, S. Integrating granger causality and vector auto-regression for traffic prediction of large-scale WLANs. *KSII Trans. Internet Inf. Syst.* **10**, 136–151. https://doi.org/10.3837/tiis.2016.01.008 (2016).
4. Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. & Vapnik, V. Support vector regression machines. In *Proceedings of the 10th International Conference on Neural Information Processing Systems*, NIPS'96, 155–161 (MIT Press, 1996).
5. Yu, B., Yin, H. & Zhu, Z. Spatio-temporal graph convolutional neural network: A deep learning framework for traffic forecasting. *CoRR***abs/1709.04875** arXiv:1709.04875 (2017).
6. Zheng, C., Fan, X., Wang, C. & Qi, J. Gman: A graph multi-attention network for traffic prediction arXiv:1911.08415 (2019).
7. Cui, Z., Chen, W. & Chen, Y. Multi-scale convolutional neural networks for time series classification arXiv:1603.06995 (2016).
8. Li, Y., Yu, R., Shahabi, C. & Liu, Y. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting arXiv:1707.01926 (2018).
9. Zhao, L. et al. T-GCN: A temporal graph convolutional network for traffic prediction. *IEEE Trans. Intell. Transp. Syst.* **21**, 3848–3858. https://doi.org/10.1109/TITS.2019.2935152 (2020).
10. Guo, S., Lin, Y., Feng, N., Song, C. & Wan, H. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 33, 922–929 (2019).
11. Wang, X. et al. Traffic flow prediction via spatial temporal graph neural network. In *Proceedings of The Web Conference 2020*, WWW '20, 1082–1092, https://doi.org/10.1145/3366423.3380186 (Association for Computing Machinery, 2020).
12. Luo, R., Song, Y., Huang, L., Zhang, Y. & Su, R. Stgin: A spatial temporal graph-informer network for long sequence traffic speed forecasting arXiv:2210.01799 (2022).
13. Wu, Z., Pan, S., Long, G., Jiang, J. & Zhang, C. Graph wavenet for deep spatial-temporal graph modeling arXiv:1906.00121 (2019).
14. Liu, Z., Shojaee, P. & Reddy, C. K. Graph-based multi-ode neural networks for spatio-temporal traffic forecasting arXiv:2305.18687 (2023).
15. Ghosh, B., Basu, B. & O'Mahony, M. Bayesian time-series model for short-term traffic flow forecasting. *J. Transp. Eng.* **133**, 180–189 (2007).
16. Xie, Y., Zhang, Y. & Ye, Z. Short-term traffic volume forecasting using Kalman filter with discrete wavelet decomposition. *Computer-Aided Civil Infrastruct. Eng.* **22**, 326–334 (2007).
17. Fu, H., Ma, H., Liu, Y. & Lu, D. A vehicle classification system based on hierarchical multi-svms in crowded traffic scenes. *Neurocomputing* **211**, 182–190 (2016).
18. May, M., Hecker, D., Körner, C., Scheider, S. & Schulz, D. A vector-geometry based spatial knn-algorithm for traffic frequency predictions. In *2008 IEEE International Conference on Data Mining Workshops*, 442–447, https://doi.org/10.1109/ICDMW.2008.35 (2008).
19. Huang, W., Song, G., Hong, H. & Xie, K. Deep architecture for traffic flow prediction: Deep belief networks with multitask learning. *IEEE Trans. Intell. Transp. Syst.* **15**, 2191–2201. https://doi.org/10.1109/TITS.2014.2311123 (2014).
20. Lv, Y., Duan, Y., Kang, W., Li, Z. & Wang, F.-Y. Traffic flow prediction with big data: A deep learning approach. *IEEE Trans. Intell. Transp. Syst.* **16**, 865–873. https://doi.org/10.1109/TITS.2014.2345663 (2015).
21. Fu, R., Zhang, Z. & Li, L. Using lstm and gru neural network methods for traffic flow prediction. In *2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, 324–328, https://doi.org/10.1109/YAC.2016.7804912 (2016).
22. Sutskever, I., Vinyals, O. & Le, Q. V. Sequence to sequence learning with neural networks arXiv:1409.3215 (2014).
23. Wang, Y., Guo, Y., Wei, Z., Huang, Y. & Liu, X. Traffic flow prediction based on deep neural networks. In *2019 International Conference on Data Mining Workshops (ICDMW)*, 210–215, https://doi.org/10.1109/ICDMW.2019.00040 (2019).
24. Wu, Y. & Tan, H. Short-term traffic flow forecasting with spatial-temporal correlation in a hybrid deep learning framework arXiv:1612.01022 (2016).
25. Yu, H., Wu, Z., Wang, S., Wang, Y. & Ma, X. Spatiotemporal recurrent convolutional networks for traffic prediction in transportation networks arXiv:1705.02699 (2017).

26. Chen, W. & Shi, K. Multi-scale attention convolutional neural network for time series classification. *Neural Netw.* **136**, 126–140. https://doi.org/10.1016/j.neunet.2021.01.001 (2021).
27. Chen, Z., Ma, Q. & Lin, Z. Time-aware multi-scale rnns for time series modeling. In *IJCAI*, 2285–2291 (2021).
28. Bai, L., Yao, L., Li, C., Wang, X. & Wang, C. Adaptive graph convolutional recurrent network for traffic forecasting arXiv:2007.02842 (2020).
29. Li, M. et al. Traffic flow prediction with vehicle trajectories. In *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 35, 294–302 (2021).
30. Wu, Z., Pan, S., Long, G., Jiang, J. & Zhang, C. Graph wavenet for deep spatial-temporal graph modeling. arXiv:1906.00121 (2019).
31. Peng, H. et al. Dynamic graph convolutional network for long-term traffic flow prediction with reinforcement learning. *Inf. Sci.* **578**, 401–416 (2021).
32. Wang, Y., Fang, S., Zhang, C., Xiang, S. & Pan, C. Tvgcn: Time-variant graph convolutional network for traffic forecasting. *Neurocomputing* **471**, 118–129 (2022).
33. Li, C., Bai, L., Liu, W., Yao, L. & Waller, S. T. A multi-task memory network with knowledge adaptation for multimodal demand forecasting. *Transp. Res. Part C Emerg. Technol.* **131**, 103352 (2021).
34. Wang, P., Zhang, T., Zheng, Y. & Hu, T. A multi-view bidirectional spatiotemporal graph network for urban traffic flow imputation. *Int. J. Geogr. Inf. Sci.* **36**, 1231–1257 (2022).
35. Huang, X., Ye, Y., Yang, X. & Xiong, L. Multi-view dynamic graph convolution neural network for traffic flow prediction. *Expert Syst. Appl.* **222**, 119779 (2023).
36. Vaswani, A. et al. Attention is all you need. arXiv:1706.03762 (2023).
37. Huang, J., Luo, K., Cao, L., Wen, Y. & Zhong, S. Learning multiaspect traffic couplings by multirelational graph attention networks for traffic prediction. *IEEE Trans. Intell. Transp. Syst.* **23**, 20681–20695. https://doi.org/10.1109/TITS.2022.3173689 (2022).
38. Feng, A. & Tassiulas, L. Adaptive graph spatial-temporal transformer network for traffic flow forecasting. arXiv:2207.05064 (2022).
39. Jiang, J., Han, C., Zhao, W. X. & Wang, J. Pdformer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction. arXiv:2301.07945 (2024).
40. Li, Y. et al. Ddgformer: Direction- and distance-aware graph transformer for traffic flow prediction. *Knowl.-Based Syst.* **302**, 112381 (2024).
41. Berndt, D. J. & Clifford, J. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, AAAIWS'94, 359–370 (AAAI Press, 1994).
42. Hamilton, J. D. *Time Series Analysis* (Princeton University Press, 1994).
43. Song, C., Lin, Y., Guo, S. & Wan, H. 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, 914–921 (2020).
44. Wu, Z. et al. Connecting the dots: Multivariate time series forecasting with graph neural networks arXiv:2005.11650 (2020).
45. Li, M. & Zhu, Z. Spatial-temporal fusion graph neural networks for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, 4189–4196 (2021).
46. Fang, Z., Long, Q., Song, G. & Xie, K. Spatial-temporal graph ode networks for traffic flow forecasting. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & amp; Data Mining*, 364–373 (ACM, 2021).
47. Lan, S. et al. Dstagnn: Dynamic spatial-temporal aware graph neural network for traffic flow forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, vol. 162 of *Proceedings of Machine Learning Research* (eds Chaudhuri, K. et al.) 11906–11917 (PMLR, 2022).
48. Xu, Y. et al. Generic dynamic graph convolutional network for traffic flow forecasting. *Inf. Fusion* **100**, 101946 (2023).
49. Liu, A. & Zhang, Y. Spatial-temporal dynamic graph convolutional network with interactive learning for traffic forecasting. *IEEE Trans. Intell. Transp. Syst.* **25**, 7645–7660. https://doi.org/10.1109/TITS.2024.3362145 (2024).
50. Guo, S., Lin, Y., Wan, H., Li, X. & Cong, G. Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting. *IEEE Trans. Knowl. Data Eng.* **34**, 5415–5428. https://doi.org/10.1109/TKDE.2021.3056502 (2022).
51. Shao, Z., Zhang, Z., Wang, F., Wei, W. & Xu, Y. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. arXiv:2208.05233 (2022).
52. Liu, H. et al. Staeformer: Spatio-temporal adaptive embedding makes vanilla transformer sota for traffic forecasting. arXiv:2308.10425 (2023).

Author contributions

Han.W. and G.T. developed the methodology; G.T. and Hao.W. were responsible for software development; Z.H., M.Z. and Z.Q. performed validation and investigation; G.T., M.Z. and Z.Q. provided resources; Z.H. handled data curation; Han.W. wrote the original draft and prepared figures; Han.W., G.T. and Z.Q. reviewed and edited the manuscript; Hao.W. was responsible for visualization; Han.W., G.T. and Z.Q. supervised the project; M.Z. managed the project administration; Z.Q. gave the funding. All authors reviewed the manuscript.

Funding

This research was funded by Project supported by the Fund for Less Developed Regions of the National Natural Science Foundation of China (Grant No. 62366044).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.