Current Science
CS 6 (4), 6387-6399 (2026)
**Received 26 May** 2026 | **Accepted 27 June** 2026 | **Online Available 01 July** 2026
https://doi.org/10.52845/CS/2026-6-4-1

**Current Science**
Current Science

**OPEN ACCESS**
**ISSN (O) 2795-8639**

**Original Article**

# A Multi-Scale Spatiotemporal Feature Self-Adaptive Fusion Model for Traffic Flow Prediction

**Chen Ning[1], Hu Chunyang[1], Zhang Songtao[2], He Linghua[2], Ma Yangyang[3], Zhang Zhiyi[1]***

**[1]School of Intelligent Manufacturing and Energy Engineering, Zhejiang University of Science and Technology,**

**Zhejiang Provincial Key Lab of Food Logistics Equipment and Technology, Hangzhou 310023, Zhejiang, China**

**[2]Zhejiang Weisong Cold Chain Technology Co., Ltd., Hangzhou, Zhejiang, China**

**[3]Jiangsu Pingjian Appraisal & Evaluation Co., Ltd., Nanjing, Jiangsu, China**

***Corresponding Author: Zhang Zhiyi**

**Abstract:**
Accurate short-term traffic flow prediction remains challenging because of the nonlinear, multi-scale and spatiotemporally coupled nature of traffic data. To address these issues, this study proposes a multi-scale spatiotemporal feature adaptive fusion model, termed MSCA-Former. The model employs a multi-branch convolution structure to capture spatial features under different receptive fields, introduces an SE-based channel recalibration mechanism to enhance informative features adaptively, and incorporates a cross-time residual enhancement module to strengthen temporal dependency modelling. In addition, an adaptive feature fusion strategy is designed to improve the integration of spatial and temporal representations. Experiments on four public datasets, namely PEMS03, PEMS04, PEMS07 and PEMS08, show that the proposed model consistently outperforms seven representative baseline models. In particular, on the PEMS08 dataset, the proposed model reduces MAE, RMSE and MAPE by 7.5%, 12.0% and 8.6%, respectively, compared with the best baseline model. On the PEMS04 dataset, the corresponding reductions in MAE and MAPE are 7.6% and 4.1%, respectively. Ablation studies further confirm that the multi-branch convolution structure, channel recalibration mechanism and cross-time residual enhancement module all contribute positively to the overall model performance. These results indicate that MSCA-Former can effectively improve the representation and prediction of complex traffic flow patterns, providing an effective approach for short-term traffic flow forecasting in intelligent transportation systems.

**Keywords: traffic flow prediction; multi-scale feature extraction; channel adaptive; time-series feature modeling**

## Introduction

Traffic flow prediction exhibits significant spatiotemporal dual characteristics, meaning that traffic conditions are not only influenced by historical temporal evolution but are also closely related to the spatial structure of the road network. Traditional traffic flow prediction methods such as ARIMA [1] and Kalman filtering [2] perform well in simple scenarios, but often struggle to model complex, nonlinear traffic data accurately. In recent years, the rapid development of deep learning technologies has provided new solutions for traffic flow prediction. In particular, convolutional neural networks (CNN) [3] and recurrent neural networks (RNN) [4] have

demonstrated powerful capabilities in handling time series data, while the Transformer model has made groundbreaking progress in capturing long-range dependencies through its self-attention mechanism. However, existing deep learning models still face numerous challenges in the application of traffic flow prediction. For example, single convolutional structures struggle to effectively extract multi-scale spatiotemporal features, the standard Transformer [5] lacks adaptive capabilities for channel features, model designs are often limited by divisibility constraints, and the spatiotemporal interaction mechanisms are not sufficiently refined. These issues constrain the further improvement of prediction performance.

Currently, traffic flow prediction research can be categorized into four types of models. The first category is CNN-based models. Xu et al. [6] proposed the RFNet model, which combines the advantages of graph convolutional networks (GCN) and deep learning, using temporal modeling and graph structure processing techniques. By designing a graph convolution-based network architecture, it effectively extracts spatial dependencies in traffic flow. Huang et al. [7] proposed the ReCoNet model, which adopts an ensemble method with multi-channel input and a convolutional neural network (CNN). This model maintains high prediction accuracy on large-scale traffic datasets while reducing computational costs and training time. The second category is RNN-based models. Zhang [8] proposed a traffic flow prediction model based on motion equations, which combines motion equations and traditional regression methods, ensuring high accuracy while further enhancing the model's flexibility and adaptability. Zhang et al. [9] proposed a multi-task automatic code generation model, which integrates parallel encoding functions for multiple tasks. By combining the use of public and private components, this model can process and optimize traffic flow prediction tasks more effectively in distributed systems. The third category is Transformer-based models. Xie et al. [10] proposed the PIAFusion model, which uses the Transformer architecture and captures spatiotemporal dependencies through its self-attention mechanism, enhancing the ability to handle complex data. Liu et al. [11] proposed the STAE-BiSSSM model, where the STAE part extracts spatiotemporal features through multi-layer embedding, improving the model's adaptability to complex spatiotemporal data. Compared to traditional Transformer models, the STAE-BiSSSM reduces the number of model parameters significantly while maintaining high performance and improving computational efficiency. The fourth category is hybrid models. Wang et al. [12] proposed a hybrid CNN-LSTM-PSO model, which first classifies traffic flow features for different stations through spectral clustering, then uses CNN to extract spatial features, LSTM to capture temporal dependencies, and optimizes the model's prediction accuracy using PSO. Zhao et al. [13] proposed the Transformer-based CDFuse network model, which demonstrates strong global dependency modeling capability in traffic flow prediction and helps overcome accuracy loss caused by the separation of spatial and temporal information in traditional methods. Zhao et al. [14] proposed a graph convolutional network model based on adaptive spatiotemporal attention, which combines long short-term memory (LSTM) and graph convolutional networks (GCN), and effectively captures the spatiotemporal features of traffic flow data using an adaptive spatiotemporal attention mechanism. This enhances the model's ability to extract complex spatiotemporal dynamics while considering the impact of road network topology on spatiotemporal correlation.

Although the above research has advanced the development of traffic flow prediction, there are still some common issues that need to be

addressed. Most models adopt a single convolutional structure, which makes it difficult to simultaneously capture local details and global trends in traffic flow data. The weight distribution of feature channels is often fixed, unable to dynamically adjust based on the input data, limiting the model's ability to focus on key information. Traditional designs often require the embed-dim to be divisible by the number of branches, a constraint that reduces the model's applicability. Furthermore, the interaction between the time dimension and the channel dimension is not deeply explored, making it difficult to fully characterize the complex dynamic evolution of traffic flow.

To address these issues, this paper proposes a multi-scale spatiotemporal feature adaptive fusion model. The model uses a multi-branch convolutional structure to extract multi-scale features under different receptive fields, introduces the SEChannelGate module for channel-level dynamic calibration, and designs a cross-time residual attention mechanism in the Transformer encoder to enhance the time-channel interaction relationship. At the same time, an automatic channel allocation mechanism is built to remove the divisibility constraint between embed-dim and the number of branches, thereby improving the model's flexibility and generalization ability.

## 2. Materials and Methods

### 2.1 Problem Definition

The traffic road network is represented as a graph $G=(V,E)$ where $V$ denotes the set of sensor nodes and $E$ denotes the spatial dependencies among nodes. Let $N$ be the number of nodes, $F$ the dimension of input features at each node, $T$ the length of the historical observation window, and $H$ the forecasting horizon. The historical traffic observations can be expressed as

$$X=\{X_{t-T+1},X_{t-T+2},\ldots,X_t\}\in \mathrm{R}^{T\times N\times F} \quad (1)$$

where $X_t \in \mathbb{R}^{N\times F}$ denotes the traffic state at time step $t$.

The objective of short-term traffic flow forecasting is to predict the traffic flow sequence over the next $H$ time steps based on the historical observations. The target sequence is defined as

$$Y=\{Y_{t+1},Y_{t+2},\ldots,Y_{t+H}\}\in \mathrm{R}^{H\times N\times C} \quad (2)$$

where $C$ denotes the output feature dimension. In this study, the forecasting task is formulated as an end-to-end supervised regression problem:

$$Y=f(X;\theta) \quad (3)$$

Where $f(\cdot)$ denotes the proposed MSCA-Former model and $\theta$ denotes the set of trainable parameters. To improve the representation ability for complex traffic states, a unified prediction framework integrating multi-scale spatial feature extraction, temporal dependency modeling, and adaptive feature fusion is constructed.

### 2.2 Overall architecture of MSCA-Former

The overall architecture of the proposed MSCA-Former is shown in Fig. 1. The model mainly consists of three components: a multi-scale spatial feature extraction module, a temporal feature modeling module, and an adaptive feature fusion and regression output module. This architecture is designed to address three major challenges in traffic flow forecasting: the insufficient representation capability of single-scale feature extraction, the limited ability of conventional temporal models to capture long-range dependencies and non-stationary dynamics, and the lack of an effective mechanism for dynamically integrating heterogeneous spatiotemporal features.

First, the historical traffic observation sequence is taken as the model input and organized in the form of "time steps × nodes × feature

dimensions". After being mapped through an embedding layer and combined with positional encoding, the input features are fed into the multi-scale spatial feature extraction module. This module employs three parallel one-dimensional convolutional branches with different kernel sizes to extract spatial features under different receptive fields, thereby enhancing the model's capability to jointly represent local fluctuations and overall trend variations in traffic flow. Subsequently, the multi-branch convolutional features are concatenated and passed into the SE channel recalibration module, which learns the importance weights of each channel to enhance key feature responses while suppressing redundant noise information, thus obtaining an enhanced spatial feature representation.

On this basis, the enhanced spatial features are input into the temporal feature modeling module. This module is built upon the Transformer encoder architecture. It employs a multi-head self-attention mechanism to capture global temporal dependencies across different time steps. Combined with a feed-forward network, residual connections, layer normalization, and a cross-temporal residual attention mechanism, it further enhances the representational capacity of temporal features, thereby obtaining the temporal branch features.

Finally, the enhanced spatial features from the spatial branch and the temporal features from the temporal branch are jointly fed into the adaptive feature fusion module, where dynamic weight allocation is used to achieve synergistic fusion of spatial and temporal information. The fused spatial-temporal features are then passed into the regression output module, ultimately producing the traffic flow prediction results for future time steps..

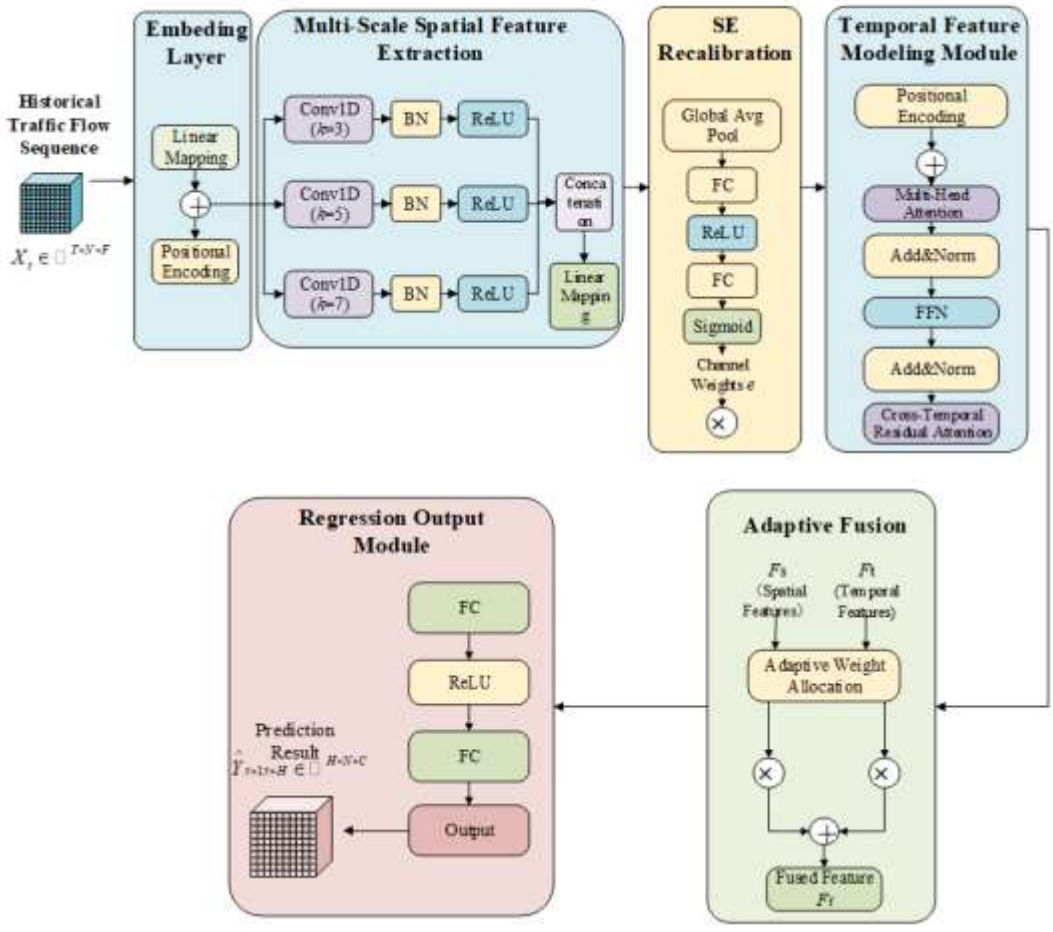



**Fig.1 : Architecture of the MSCA-Former model**

### 2.3 Multi-Scale Spatial Feature Extraction Module

Traffic flow data usually exhibit significant spatial heterogeneity and multi-scale variation characteristics. In real traffic systems, local neighborhood interactions, medium-range propagation effects, and broader traffic evolution trends may coexist. If only a single-scale convolution kernel is adopted, it is difficult to simultaneously capture local details and global contextual information. Therefore, a multi-branch convolution structure is designed in this study to enhance the model's ability to perceive spatial features at different scales.

For the input sequence $X$, three parallel convolution branches are constructed with kernel sizes of 3, 5, and 7, respectively. The branch with kernel size 3 is used to capture fine-grained local variation patterns, the branch with kernel size 5 is used to model medium-range spatial patterns, and the branch with kernel size 7 is used to perceive broader contextual information. For the $i$-th branch, the feature extraction process can be written as

$$F_i = \sigma(W_i * X + b_i), \quad i \in \{1,2,3\} \qquad (4)$$

where $W_i$ and $b_i$ denote the convolution kernel and bias of the $i$-th branch, respectively, $*$ denotes the convolution operation, and $\sigma(\cdot)$ denotes the nonlinear activation function.

After obtaining the outputs of all branches, the extracted features at different scales are concatenated along the channel dimension to form the initial multi-scale spatial representation:

$$F_{ms} = \mathrm{Concat}(F_1, F_2, F_3) \qquad (5)$$

This structure enables the model to characterize spatial dependencies in traffic flow from multiple receptive fields and reduces the risk of information loss caused by single-scale modeling.

To further improve the discriminability of multi-scale features, an SE-based channel recalibration mechanism is introduced after multi-branch convolution. Specifically, global average pooling is first applied to the multi-scale feature map $F_{ms}$, compressing the spatial dimensions into a channel descriptor vector:

$$z_c = \frac{1}{H_s W_s} \sum_{i=1}^{H_s} \sum_{j=1}^{W_s} F_{ms}^{(c)}(i,j) \qquad (6)$$

where $z_c$ denotes the aggregated descriptor of the $c$-th channel, and $H_s$ and $W_s$ denote the spatial dimensions of the feature map.

Then, two fully connected layers are used to model channel-wise dependencies and generate channel weights:

$$s = \sigma\left(W_2 \delta(W_1 z)\right) \qquad (7)$$

where $W_1$ and $W_2$ are learnable parameter matrices, $\delta(\cdot)$ denotes the ReLU activation function, and $\sigma(\cdot)$ denotes the Sigmoid activation function.

Finally, the channel weights are multiplied with the multi-scale features to obtain the enhanced spatial representation:

$$\hat{F}_{ms} = s \odot F_{ms} \qquad (8)$$

where $\odot$ denotes element-wise multiplication. Through this mechanism, the model can adaptively emphasize feature channels that contribute more to traffic prediction while suppressing noisy or redundant responses.

In addition, to improve the flexibility of the multi-branch structure, an adaptive channel allocation strategy is adopted, which avoids the strict divisibility constraint between the embedding dimension and the number of branches in conventional multi-branch designs, thereby enhancing architectural flexibility and adaptability.

### 2.4 Temporal feature modeling module

Although the multi-scale convolution structure can effectively extract spatial representations, traffic flow forecasting is essentially a temporal modeling

task. Traffic states are jointly influenced by recent observations, periodic fluctuations, long-term trends, and unexpected disturbances. Therefore, a Transformer-based temporal encoder is employed to model both short-term and long-term dependencies in traffic flow sequences.

Given the spatially enhanced

feature sequence $\hat{F}_{ms}$, it is first reorganized into a

temporal representation suitable for sequence modeling and then projected into the query, key, and value matrices:

$$Q = \hat{F}_{ms} W_Q, \quad K = \hat{F}_{ms} W_K, \quad V = \hat{F}_{ms} W_V \tag{9}$$

where $W_Q$, $W_K$, and $W_V$ are learnable projection matrices.

The scaled dot-product attention is then applied to compute temporal dependencies across different time steps:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V \tag{10}$$

where $d_k$ denotes the dimension of the key vector. This mechanism directly establishes dependencies between arbitrary time steps and effectively alleviates the information attenuation problem commonly observed in recurrent neural networks when modeling long sequences.

To further enhance representation capability, a multi-head attention mechanism is employed:

$$\text{MultiHead}(Q,K,V) = \text{Concat}(head_1,\ldots,head_h)W_O \tag{11}$$

where $head_i$ denotes the output of the $i$-th attention head, $h$ denotes the number of attention heads, and $W_O$ denotes the output projection matrix. The multi-head mechanism enables the model to learn diverse temporal dependencies from different subspaces.

After the multi-head attention layer, residual connection and layer normalization are applied to improve training stability and mitigate optimization difficulty in deep architectures:

$$Z^{(1)} = \text{LayerNorm}\left(\hat{F}_{ms} + \text{MultiHead}(Q,K,V)\right) \tag{12}$$

The resulting feature is then fed into a feed-forward network to further extract high-level temporal representations, followed by another residual connection and layer normalization:

$$Z^{(2)} = \text{LayerNorm}\left(Z^{(1)} + \text{FFN}(Z^{(1)})\right) \tag{13}$$

Through the above process, the model can capture both short-term fluctuation patterns and long-range temporal structures in traffic flow sequences.

### 2.4.1 Cross-Time Residual Enhancement Mechanism

Although the standard Transformer has strong global dependency modeling ability, it mainly relies on similarity-based attention computation. For traffic flow data, abrupt changes, non-stationary fluctuations, and strong couplings across distant time steps are common, and such patterns may not be sufficiently captured by standard self-attention alone. Therefore, a cross-time residual enhancement mechanism is further introduced to strengthen temporal interactions across different time steps.

Let $Z^{(2)}$ denote the temporal feature generated by the Transformer encoder. The cross-time residual enhancement process can be expressed as

$$\tilde{Z} = Z^{(2)} + \phi(Z^{(2)}) \tag{14}$$

where $\phi(\cdot)$ denotes the cross-time residual mapping function, which can be implemented by a lightweight linear transformation, a residual attention mapping, or another enhancement operation along the temporal dimension. By introducing an additional temporal information propagation path on top of the original representation, this mechanism can further strengthen the coupling relationship among critical time steps.

By combining the standard Transformer encoder with the cross-time residual enhancement

mechanism, the model can not only capture global temporal dependencies in traffic flow sequences but also become more sensitive to abrupt variations, non-stationary characteristics, and complex dynamic changes, thereby improving forecasting stability and accuracy.

### 2.5 Adaptive Feature Fusion and Regression Output Module

After spatial feature extraction and temporal dependency modeling, how to effectively integrate these two heterogeneous representations becomes a key issue affecting prediction performance. If simple addition or fixed-weight fusion is adopted, the complementary advantages of spatial and temporal information may not be fully exploited, and the model may not adapt well to dynamic changes in feature importance under different traffic conditions. Therefore, an adaptive feature fusion module is designed to dynamically allocate weights to different feature sources and achieve collaborative modeling.

Let $F_s$ denote the feature representation extracted from the spatial branch and $F_t$ denote the feature representation extracted from the temporal branch. To improve feature compatibility, they are first projected into a unified latent space:

$$\bar{F}_s = W_s F_s + b_s, \quad \bar{F}_t = W_t F_t + b_t \qquad (15)$$

where $W_s$, $W_t$, $b_s$, and $b_t$ are learnable parameters.

Then, the projected spatial and temporal features are concatenated and fed into a lightweight multilayer perceptron to generate adaptive fusion weights:

$$\alpha, \beta = \mathrm{Softmax}\left(\mathrm{MLP}([\bar{F}_s; \bar{F}_t])\right) \qquad (16)$$

where $[\cdot,\cdot]$ denotes feature concatenation, and $\alpha$ and $\beta$ denote the normalized weights for spatial and temporal features, respectively. These weights are dynamically learned from the input samples and can reflect the relative importance of the two feature types under different traffic states.

Based on the learned weights, the fused representation is obtained as

$$F_{fuse} = \alpha \cdot \bar{F}_s + \beta \cdot \bar{F}_t \qquad (17)$$

This formulation allows the model to adaptively adjust the contribution ratio of spatial and temporal information according to traffic dynamics, thereby improving the flexibility and robustness of feature integration.

Finally, the fused feature is fed into the regression head to generate the future traffic flow prediction:

$$\hat{Y} = W_r F_{fuse} + b_r \qquad (18)$$

where $W_r$ and $b_r$ are the parameters of the regression layer, and $\hat{Y}$ denotes the predicted traffic flow sequence over the forecasting horizon.

To further enhance optimization stability, a residual output strategy can also be introduced in the regression stage:

$$\hat{Y} = W_r F_{fuse} + b_r + R(X) \qquad (19)$$

where $R(X)$ denotes a residual mapping from the original input sequence or intermediate features. This design helps preserve low-level dynamic information and improves the fitting ability of the model for complex traffic patterns.

## 3. Results

### 3.1 Datasets

To evaluate the prediction performance and generalization ability of the proposed model under different traffic scenarios, four publicly available real-world traffic flow datasets, namely PEMS03, PEMS04, PEMS07, and PEMS08, are employed in this study. These datasets are collected from the Performance Measurement System (PEMS)[15] of the California highway traffic monitoring network and have been widely used in traffic flow forecasting research due to

their authenticity and representativeness.

The four datasets differ in the number of sensor nodes, sample size, and temporal coverage, which makes them suitable for evaluating the adaptability of the proposed model under traffic networks of different scales and complexities. As summarized in Table 1.

**Table 1： Statistics of the datasets**

| Dataset | Number of nodes | Number of samples | Time span |
|---|---|---|---|
| PEMS03 | 358 | 260208 | Sep.2018 to Nov.2018 |
| PEMS04 | 307 | 16992 | Jan 2018 to Feb.2018 |
| PEMS07 | 883 | 28224 | May 2018 to Aug.2018 |
| PEMS08 | 170 | 17856 | Jul.2017 to Aug.2017 |

In this study, the time interval of all datasets is unified as 5 min, which means that each hour contains 12 sampling points. The spatial graph of each dataset is constructed according to the actual road topology. During data preprocessing, linear interpolation is first adopted to fill missing values so as to reduce the influence of incomplete observations on model training and prediction. Then, Z-score normalization is applied to eliminate the scale differences among different nodes and datasets. Finally, each dataset is divided into the training set, validation set, and test set according to a ratio of 6:2:2. By using a unified preprocessing and partition strategy, the comparability of experimental results across different datasets and models can be ensured.

### 3.2 Experimental Settings

All experiments were conducted under a unified environment. The experimental platform was based on the Linux operating system, and an NVIDIA GeForce RTX 3090 Ti GPU was used for model training and testing. The proposed model was implemented using the PyTorch deep learning framework, with PyTorch version 1.12.1 and CUDA version 11.6.

In the experiments, both the input sequence length and the prediction horizon were set to 12, meaning that the traffic flow of the previous 1 h was used to predict the traffic flow of the next 1 h. The multi-scale spatial feature extraction module consisted of three parallel convolution branches with kernel sizes of 3, 5, and 7, respectively. The number of output channels was set to 128, and ReLU was adopted as the activation function. In the SE channel attention module, the channel reduction ratio was set to 16. In the temporal modeling module, the number of Transformer encoder layers was set to 2, the number of attention heads was set to 8, the embedding dimension was set to 128, the hidden dimension of the feed-forward network was set to 256, and the dropout rate was set to 0.1.

During training, the Adam optimizer was adopted, with a batch size of 64, an initial learning rate of 0.001, a weight decay coefficient of $1 \times 10^{-5}$, and a maximum of 100 training epochs. To improve training stability, the ReduceLROnPlateau strategy was employed together with an early stopping mechanism. When the validation loss did not decrease significantly for five consecutive epochs, the learning rate was reduced to half of its current value, and the early stopping patience was set to 10. In addition, the gradient clipping threshold was set to 5.0 to alleviate gradient explosion during training. To reduce the influence of randomness, the random seed was fixed at 42 for all experiments.

### 3.3 Comparison Experiments and Analysis

To comprehensively evaluate the predictive performance of the proposed MSCA-Former, comparative experiments were conducted against seven representative baseline models, including

statistical learning models, machine learning methods, recurrent neural networks, and spatiotemporal graph-based deep learning models. The selected baseline models are described as follows:

ARIMA[16]: Autoregressive integrated moving average (ARIMA) is a classical time-series analysis model that characterizes the linear variation patterns of a sequence through autoregressive, differencing, and moving average operations. It is widely used for forecasting stationary or approximately stationary time series, such as short-term traffic flow data.

SVR[17]: Support vector regression (SVR) is a supervised learning method derived from support vector machines. By mapping input data into a high-dimensional feature space through kernel functions and constructing a regression model under the principle of structural risk minimization, it is effective for forecasting nonlinear time-series data, especially in small-sample settings.

LSTM[18]: Long short-term memory (LSTM) is an improved recurrent neural network architecture that introduces gating mechanisms into traditional recurrent neural networks. It alleviates the problems of gradient vanishing and gradient explosion in long-sequence training, thereby enhancing the capability to capture long-term dependencies in time-series data.

PDFormer[19]: PDFormer is a Transformer-based spatiotemporal forecasting model that explores latent spatiotemporal dependencies and dynamic variation patterns in traffic flow data through a pattern discovery mechanism, thereby improving the accuracy and robustness of multi-step forecasting under complex traffic conditions.

DCRNN[20]: Diffusion convolutional recurrent neural network (DCRNN) combines diffusion graph convolution with recurrent neural networks. It captures both the diffusion propagation relationships among traffic nodes on directed graphs and the temporal dynamic dependencies of traffic flow, making it suitable for spatiotemporal traffic forecasting tasks.

STGCN[21]: Spatiotemporal graph convolutional network (STGCN) is a deep learning model that integrates graph convolution with temporal convolution. By jointly modeling the spatial topological correlations of traffic networks and the temporal evolution characteristics of traffic flow, it effectively captures spatiotemporal dependencies in traffic data.

ASTGCN[22]: Attention-based spatiotemporal graph convolutional network (ASTGCN) further incorporates an attention mechanism into STGCN to capture dynamic spatial and temporal patterns, thereby improving the representation of complex spatiotemporal dependencies.

Table 2 and Table 3 present the prediction results of MSCA-Former and seven baseline models on the PEMS03, PEMS04, PEMS07 and PEMS08 datasets. Overall, the proposed MSCA-Former achieves the best performance on all four datasets, indicating its strong predictive capability and good cross-dataset generalization.

Specifically, on the PEMS03 dataset, the MAE, RMSE and MAPE of MSCA-Former are 15.46, 27.08 and 15.06%, respectively; on the PEMS04 dataset, they are 18.32, 30.09 and 12.27%, respectively; on the PEMS07 dataset, they are 19.83, 32.73 and 8.48%, respectively; and on the PEMS08 dataset, they are 13.78, 22.93 and 9.28%, respectively. These results are consistently better than those of the baseline models. In particular, on the PEMS04 and PEMS08 datasets, the proposed model shows more evident improvements over the best baseline models, indicating that it can capture the spatiotemporal dependencies in traffic flow data more effectively.

As shown in Figure 2, the error metrics of all models generally increase with the prediction horizon. However, MSCA-Former consistently

maintains lower error values and a relatively smaller growth trend, demonstrating better stability and robustness in multi-step forecasting tasks.

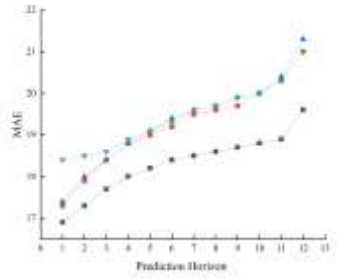
Overall, MSCA-Former shows superior predictive performance under different datasets and forecasting horizons, which verifies its effectiveness in short-term traffic flow prediction.

**Table 2. Comparison of prediction results on the PEMS03 and PEMS04 datasets**

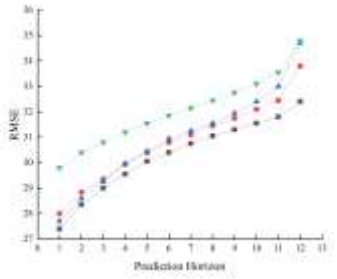
| Model | PMS03 | | | PEMS04 | | |
|---|---|---|---|---|---|---|
| | MAE | RMSE | MAPE | MAE | RMSE | MAPE |
| ARIMA | 34.41 | 46.68 | 33.76% | 33.65 | 46.70 | 24.18% |
| SVR | 19.12 | 32.77 | 18.89% | 24.89 | 39.66 | 17.19% |
| LSTM | 20.55 | 35.11 | 24.96% | 26.77 | 40.64 | 18.23% |
| PDFormer | 15.98 | 28.25 | 15.23% | 19.83 | 30.16 | 13.97% |
| DCRNN | 17.99 | 30.31 | 18.34% | 21.22 | 33.26 | 14.17% |
| STGCN | 17.48 | 29.21 | 15.53% | 21.19 | 33.65 | 12.80% |
| ASTGCN | 17.34 | 29.56 | 17.21% | 22.92 | 35.32 | 16.56% |
| **Ours** | **15.46** | **27.08** | **15.06%** | **18.32** | **30.09** | **12.27%** |

**Table 3. Comparison of experimental results on PEMS07 and PEMS08**

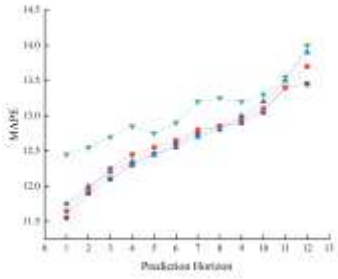
| Model | PEMS07 | | | PEMS08 | | |
|---|---|---|---|---|---|---|
| | MAE | RMSE | MAPE | MAE | RMSE | MAPE |
| ARIMA | 37.56 | 58.41 | 19.46% | 32.31 | 44.36 | 22.73% |
| SVR | 22.65 | 41.50 | 18.18% | 18.27 | 31.20 | 12.13% |
| LSTM | 28.89 | 46.49 | 13.21% | 19.18 | 26.91 | 13.21% |
| PDFormer | 22.37 | 36.55 | 9.23% | 14.89 | 26.21 | 10.15% |
| DCRNN | 25.22 | 38.61 | 11.91% | 16.78 | 26.36 | 10.92% |
| STGCN | 25.33 | 39.34 | 11.21% | 17.40 | 27.09 | 11.29% |
| ASTGCN | 24.01 | 36.67 | 10.73% | 18.25 | 26.06 | 11.64% |
| **Ours** | **19.83** | **32.73** | **8.48%** | **13.78** | **22.93** | **9.28%** |



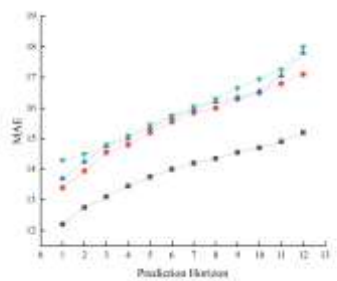
(a) MAE on PEMS04



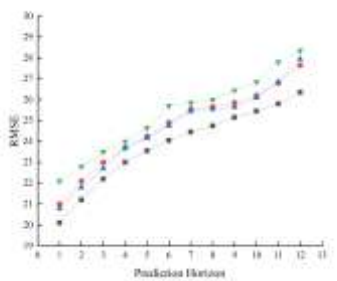
(b) RMSE on PEMS04



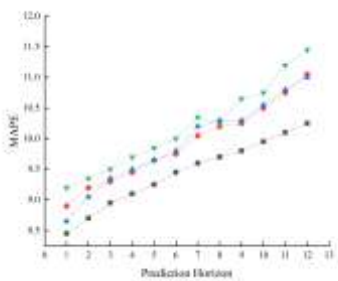
(c) MAPE on PEMS04



(d) MAE on PEMS04



(e) RMSE on PEMS04



(f) MAPE on PEMS04

**Fig. 2: Comparison of MAE, RMSE and MAPE of different models under different prediction horizons on the PEMS04 (top) and PEMS08 (bottom) datasets**

## 4. Discussion

### 4.1 Ablation study

To further examine the rationality and necessity of different structural components in the proposed model, and to verify the actual contribution of each module during feature modelling, ablation experiments were conducted on the PEMS04 and PEMS08 datasets under the same dataset setting, training strategy and parameter configuration. Specifically, a series of ablation studies was designed in which key modules of the model were removed or simplified one by one, and the predictive performance before and after ablation was compared, so as to systematically evaluate the effect of each component on the overall model performance. The experimental results are presented in Table 4, and the three model variants are described as follows:

Ablation study on the multi-scale convolution module (No-Multibranch): The multi-branch temporal convolution structure was removed, and only a single-scale convolution branch was retained for feature extraction.

Ablation study on the channel-adaptive recalibration module (No-SE): The SE module was removed from the model to evaluate its effect on channel feature recalibration.

Ablation study on the cross-time residual attention module (No-Cross-Attention): The cross-time residual attention module was removed, while the Transformer encoder structure was retained, in order to assess the role of cross-time interaction modelling.

The experimental results show that the complete MSCA-Former achieves better MAE, RMSE and MAPE values than all ablation variants on both the PEMS04 and PEMS08 datasets, indicating that the collaboration of the key modules can effectively improve the prediction performance.

On the PEMS04 dataset, the most evident performance degradation is observed after removing the multi-branch convolution structure, with MAE, RMSE and MAPE increasing to 19.42, 30.93 and 13.14%, respectively. This indicates that the multi-branch convolution plays an important role in extracting multi-scale spatial features and provides the basis for modelling complex spatial dependencies in traffic flow. After removing the SE channel attention mechanism, all evaluation metrics also increase, suggesting that this module can enhance the representation of important features and suppress redundant information through channel-wise weighting. Similarly, removing the cross-time residual enhancement module leads to performance degradation, showing that this module helps strengthen the modelling of temporal dependencies across different time steps and improves temporal feature representation.

The results on the PEMS08 dataset are generally consistent with those on PEMS04. The complete model still achieves the best performance, while removing the multi-branch convolution results in the most obvious decline, further confirming the importance of multi-scale spatial feature extraction. In addition, removing the SE module or the cross-time residual enhancement module also leads to increased prediction errors, indicating that both feature enhancement and temporal modelling modules are essential for improving the stability and generalization ability of the model.

**Table.4 Ablation study results**

| Model | PEMS04 | PEMS08 |
|---|---|---|

| | MAE | RMSE | MAPE | MAE | RMSE | MAPE |
|---|---|---|---|---|---|---|
| w/o Multi-branch Conv | 19.42 | 30.93 | 13.14% | 15.44 | 25.02 | 10.58% |
| w/o SE | 18.61 | 30.71 | 12.45% | 14.07 | 24.23 | 9.03% |
| w/o CTRE | 18.66 | 30.70 | 12.91% | 14.05 | 24.09 | 9.46% |
| MSCA-Former | 18.32 | 30.09 | 12.27% | 13.90 | 23.65 | 9.29% |

## 5 Conclusions

A multi-scale spatiotemporal feature adaptive fusion model, MSCA-Former, is proposed in this study for short-term traffic flow prediction. By integrating multi-branch convolution, channel recalibration and cross-time residual enhancement within a unified framework, the model is able to capture complex spatial patterns and temporal dependencies more effectively.

Experiments on four public datasets, namely PEMS03, PEMS04, PEMS07 and PEMS08, show that the proposed model consistently achieves better performance than seven representative baseline models. The results further indicate that MSCA-Former maintains good stability and robustness under different forecasting horizons.

The ablation results confirm that each proposed component contributes to the overall performance improvement. In particular, the multi-branch convolution structure, SE-based channel recalibration and cross-time residual enhancement module all play important roles in improving prediction accuracy.

Despite these promising results, the present study does not explicitly incorporate external factors such as weather, incidents or special events, and the computational cost of the model still requires further optimization for real-time applications. Future work will therefore focus on integrating heterogeneous traffic-related information, improving computational efficiency, and extending the proposed framework to more complex intelligent transportation scenarios.

**References**:

1. Hai L, Liu W, Liu Y, et al. Metro passenger flow prediction based on the ARIMA algorithm[J]. Computer and Digital Engineering, 2025(3).
2. Shen C, Wang G, Liu L, et al. Clock bias prediction algorithm based on Kalman filtering corrected by frequency difference estimation[J]. Acta Geodaetica et Cartographica Sinica, 2025, 54(9): 1596-1607.
3. Dofitas C., Gil J.-M., Byun Y.-C. Multi-directional long-term recurrent convolutional network for road situation recognition[J]. Sensors, 2024, 24(4618)..
4. Liu Z, Li X, Lu Z, et al. IWOA-RNN: An improved whale optimization algorithm with recurrent neural networks for traffic flow prediction[J]. Alexandria Engineering Journal , 2025, 117: 563–576.
5. Li W, Chen J, Zhang Y, et al. MSGFormer: Revolutionizing Traffic Flow Prediction with Multiscale and Gated Transformer Architecture[J]. IEEE Internet of Things Journal, 2025, 12(2):2014-2025.
6. Li, F, Feng, J, Yan, H, et al. Dynamic graph convolutional recurrent network for traffic prediction: Benchmark and solution. ACM Transactions on Knowledge Discovery from Data, 2023,17(1), 9.
7. Wang C, Zhang K, Wang H, et al. Auto-STGCN: Autonomous Spatial-Temporal Graph Convolutional Network Search [J]. ACM Transactions on Knowledge Discovery from Data, 2023, 17(5):21.
8. Zhang X. Research on short-term traffic flow prediction based on K-means and Bi-LSTM [J]. China Storage & Transport, 2024(11): 64-65.
9. Zhang X, Su J, Chen X, et al. Short-term traffic flow prediction based on Attention-T-

GRU [J]. Journal of Lanzhou University of Technology, 2025, 51(4): 107-113.
10. Xie G, Zhang W, He L, et al. Building damage assessment model based on UCT ransNet[J]. Computer Engineering and Design, 2025, 46(1): 44-51.
11. Liu D, Qu Q, Chen X. STAE-BiSSSM: A traffic flow forecasting model with high parameter effectiveness[J]. Isprs International Journal of Geo-Information, 2025, 14: 388.
12. Wang T, Peng K, Xiao D, et al. Integrating spectral clustering and hybrid CNN-LSTM-PSO model for short-term passenger flow prediction in urban rail transit[J]. IET Intelligent Transport Systems, 2025, 19:e 70073.
13. Du X, Li H. ENSO prediction model based on an integrated GCN-Transformer network [J]. Acta Oceanologica Sinica, 2023 (12): 45.
14. Xiao H, Zou B, Xiao J. Graph convolution networks based on adaptive spatiotemporal attention for traffic flow forecasting[J]. Scientific Reports, 2025, 15:8935.
15. CHEN C, PETTY K, SKABARDONIS A,et al. Freeway performance measurement system: mining loop detector data[J]. Transportation Research Record, 2001, 1748 (1): 96- 102.
16. Hamilton Jd. Time series analysis[M]. Princeton: Princeton University Press, 1994.
17. Castro-Neto M, Jeong Ys, Jeong Mk, et al. Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions[J]. Expert Systems with Applications, 2009, 36(3): 6164-6173.
18. Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural Computation, 1997, 9 (8): 1735-1780.
19. Wang Jy, Du H, Zeng Yy. PDFormer: Efficient vision transformer for photovoltaic defect detection[J]. IEEE Transactions on Consumer Electronics, 2025,71(2):6602-661 1.
20. Li Y, Yu R, Shahabi C, et al. Diffusion convolutional recurrent neural network: data-driven traffic forecasting[EB/OL]. arXiv:17 07.01926,2018-02-22[2025-03-28]. https:// arxiv.org/abs/1707.01926.
21. Yu B, Yin H, Zhu Z. Spatio-Temporal graph convolutional networks: a deep learning framework for traffic forecasting[C]// Proceedings of the International Joint Conference on Artificial Intelligence, 2018: 3634-3640.
22. Guo S, Lin Y, Feng N, et al. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting[C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 922-929.