Citation: Leyao Xiao, Qian Chen. multi-step short-term traffic flow prediction of urban road network based on ISTA-transformer model. *Journal of Harbin Institute of Technology(New Series)*. DOI:10.11916/j.issn.1005-9113.24075

Multi-step Short-term Traffic Flow Prediction of Urban Road Network Based on ISTA-Transformer Model

Leyao Xiao and Qian Chen *

(*School of Transportation, Southeast University, Nanjing 211189, China*)

Abstract: Short-term traffic flow prediction plays a crucial role in planning of intelligent transportation systems. Nowadays, there is a large amount of traffic flow data generated from the monitoring devices of urban road networks, which contains road network traffic information with high application value. In this study, an improved spatio-temporal attention transformer model (ISTA-Transformer Model) is proposed to provide a more accurate method for predicting multi-step short-term traffic flow based on monitoring data. By embedding a temporal attention layer and a spatial attention layer in the model, the model learns the relationship between traffic flows at different time intervals and different geographic locations, and realizes more accurate multi-step short-time flow prediction. Finally, we validate the superiority of the model with monitoring data spanning 15 days from 620 monitoring points in Qingdao, China. In the four time steps of prediction, the MAPE (Mean Absolute Percentage Error) values of ISTA-Transformer's prediction results are 0.22, 0.29, 0.37, and 0.38, respectively, and its prediction accuracy is usually better than that of six baseline models (Transformer, GRU, CNN, LSTM, Seq2Seq and LightGBM), which indicates that the proposed model in this paper always has a better ability to explain the prediction results with the time steps in the multi-step prediction.

Keywords: urban road network, traffic flow prediction, spatio-temporal feature, ISTA-Transformer Model

CLC number: U12 **Document code**: A **Article ID**: 1005-9113(2025)00-0000-14

0 Introduction

Intelligent Transportation Systems (ITS) have become a focal point of research in recent years, serving as a cornerstone for the development of smart cities. Traffic flow prediction, as a key component of ITS, provides critical data that enables optimised allocation of urban traffic resources, helps city managers respond quickly to varying traffic conditions, and provides travellers with more efficient route planning. However, with the continued expansion of urban road networks and the increase in vehicle ownership, the stochastic and uncertain nature of traffic flows has become more pronounced, posing significant challenges to accurate short-term forecasting within specific urban areas.

Short-term, multi-step traffic flow forecasting has significant value for traffic management in these areas, where multi-step refers to predicting a sequence of future time steps based on historical data, rather than making independent forecasts for each time step. Accurate multi-step forecasting can help urban transport authorities allocate resources in advance, manage congestion at a micro level and implement demand-based control measures tailored to specific network segments. By accurately predicting traffic flow over multiple time steps, decision makers can better deal with the unique challenges posed by sudden fluctuations, improving the accuracy and efficiency of real-time traffic management. For instance, shorter time steps (15 – 30 min) can support immediate congestion mitigation, whereas longer time steps (45–60 min) can aid in strategic planning for traffic signal adjustments and rerouting.

Although recent advances in predictive models, including deep learning and spatio-temporal methods, have improved prediction accuracy, current approaches often lack the ability to fully capture the spatio-temporal relationships between multiple monitoring points within a region. This limitation affects the robustness of predictions, especially in

Received 2024-12-12.

Sponsored by National Key Research and Development Program of China(Grant No.2020YEB1600500).

* Corresponding author.Qian Chen, Ph.D, Associate Professor.Email.101010372@seu.edu.cn.

dynamically changing urban environments. To address these challenges, this study proposes a model that exploits the spatio-temporal relationships between different monitoring points within a given area to improve the accuracy of short-term, multi-step forecasts over the next few intervals, thereby supporting more effective and timely urban traffic management.

1 Related Work

Traffic flow prediction is a typical regression problem which is about forcasting future traffic flow based on historical information[1]. As research in the field of traffic flow prediction continues to expand, a multitude of models and methodologies have been proposed. Prediction models that are most commonly utilized can be broadly classified into two categories: linear theoretical models and neural network models[2].

1.1 Linear Theoretical Models

The linear theory forecasting model incorporates traditional time series analysis methodologies, including the ARIMA (Autoregressive Integrated Moving Average) model, exponential smoothing method, the Markov Chain (MC), and so on. The ARIMA[3] method was initially proposed by Ahmed et al. in 1996, with van der Voort et al.[4] subsequently proposing a hybrid method, the Kohonen-ARIMA model, based on existing studies. Later, Lee et al.[5] used a subset ARIMA model, and Williams[6] tried using an ARIMAX model with explanatory variables. Kamarianakis et al.[7] proposed a space-time ARIMA model, while Williams et al.[8] put forward the seasonal ARIMA model by considering univariate modelling as a periodic process. Exponential smoothing is a process of repeatedly updating forecasts based on recent experience. Typically, in metropolitan areas where weekday traffic flow patterns differ from weekend traffic flow patterns, Holt-Winters exponential smoothing works well when the data is both trend and seasonality[9]. In traffic flow forecasting, triple exponential smoothing method is commonly used to display trend and seasonal data. Markov model is a powerful tool for measuring state space and analyzing time series data.

It is suitable for stochastic systems with large randomness and obvious data fluctuation[10]. These models are typically predicated on the assumption of a linear relationship or a simple change rule of data, which is applicable to relatively stable and linear relationship strong time series data forecasting[11]. Nevertheless, time series models analogous to ARIMA, despite their efficacy in linear analysis, are incapable of addressing nonlinear relationships in traffic flow due to their inadequate encapsulation of the dependencies between historical data.

1.2 Neural Network Models

Neural network prediction models include feed-forward neural networks, recurrent neural networks (such as long short-term memory networks, LSTM), convolutional neural networks (CNN), and other deep learning models. These models are capable of handling complex non-linear relationships and time series data, and enhance the accuracy and generalization ability of time series forecasting by learning patterns and features from a substantial quantity of data. The advent of artificial neural networks has opened up new avenues for research in the field of traffic flow prediction. In 2015, Ref.[12] demonstrated the applicability of deep learning methods in the field of traffic flow prediction. The application of deep learning, an extension of multi-layer artificial neural networks, has led to a significant advancement in the field of traffic flow prediction. By leveraging large amounts of historical data, deep learning has enabled a more accurate and comprehensive approach to traffic flow prediction, marking a notable shift in the research landscape. In the present era, predictive models based on deep learning networks have become the dominant paradigm in the domain of traffic flow prediction. In recent years, a significant number of scholars have sought to enhance the performance of deep learning models by improving the original models or combining different deep learning methods in order to leverage their respective advantages. For example, Ref.[13] introduced the SW-BiLSTM model, which considers the interaction between adjacent road segments and achieved impressive predictive accuracy when validated with real-world GPS data. In Ref.[14], a BiGRU-BiGRU model with two modules is proposed to extract temporal and periodic features from traffic data, where a novel limited attention mechanism is incorporated in the first module to improve the model's performance by focusing only on the most recent relevant information in the traffic flow sequence. Luo et al.[15] proposed a novel model, the

graph temporal convolutional long short-term memory network (GT-LSTM), which is primarily constituted of features splicing and patterns capturing. In features splicing, the spatial dependencies of traffic flow are captured through the employment of a self-adaptive graph convolutional network (GCN). Zhang et al.[16] proposed a deep learning framework that utilizes Seq2Seq models and graph convolutional networks to capture spatial and temporal dependencies, incorporating attention mechanisms and a novel training method to tackle multi-step prediction challenges and effectively address the temporal heterogeneity of traffic patterns.

Summarizing the literature on these different approaches, it is easy to see that most studies optimize the ability of the model to capture the temporal and spatial relationships of historical traffic flows, thus improving the accuracy of the predictions[17-20]. However, if the prediction sequences are too long, memory degradation and information loss may even occur due to the limitations of the above networks[21]. The transformer network is a sequence-to-sequence learning model that fully relies on the self-attention mechanism first used in natural language translation to compute the mapping relationship between input and output in parallel[22]. In 2017, the transformer architecture consisting entirely of attention mechanisms was introduced[22]. In recent years, it has been applied to traffic flow prediction. For example, Xiao and Chen [23] developed a novel temporal attention module in a spatio-temporal transformer graph network that uses local context to improve the stability of long-term predictions in the temporal dimension. Hu et al.[24] proposed a multi-layer model based on transformer and deep learning which uses multiple encoders and decoders to perform feature extraction on the initial traffic data without human experience. Conversely, these models only consider single-step prediction, which results in a limited total time span of prediction when the time granularity is fine, or the loss of some detailed information when the time granularity is too large. In contrast, multi-step short-time prediction is capable of further dividing the long time span and obtaining more detailed and complete prediction information[25-26].

In conclusion, this paper puts forth a multi-step prediction model to address the challenge of short-term flow forecasting at pivotal monitoring locations within an urban road network. This model, designated as the ISTA-Transformer (Improved Spatio-temporal Attention Transformer), employs a novel approach that incorporates enhanced spatial and temporal attention mechanisms. This study proposes a systematic processing framework (as shown in Fig. 1) for the collected road network monitoring point data, and optimizes the underlying transformer algorithm by innovatively embedding a spatio-temporal attention mechanism that takes the historical flow and spatial weight matrices as inputs and allows the model to capture the spatio-temporal relationships therein and predict the future short-term flow on a given road section. The process of short term flow prediction for road network monitoring point flow data can be simply divided into the following steps:

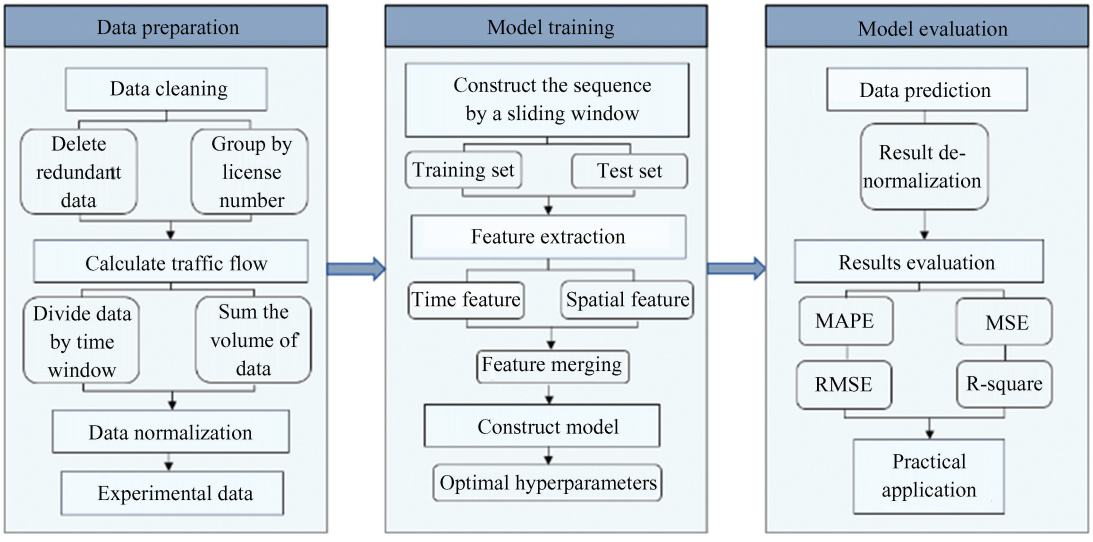


Fig. 1 Flowchart of traffic flow prediction

1) Data preparation, where toll booth traffic data is segmented by time periods, and the flow for each period is calculated, converting the data into a format that the model can process. 2) Model training, historical traffic data is used to train the constructed traffic prediction model, allowing the model to extract the spatio-temporal features contained in the data. 3) Model evaluation, predicted traffic is compared with actual traffic, and various metrics are used to assess the prediction accuracy.

2 Problem Description and Modeling

2.1 Problem Description

The short-term traffic flow multi-step prediction problem can be formulated as follows: a data series consisting of traffic flow observations from several historical periods is input into a prediction model, which outputs a prediction of a traffic flow sequence for several future time periods[27]. The objective of this study is to predict the traffic flow at various monitoring points along a road network for a number of future time steps. This is based on historical monitoring point flow data and the latitude and longitude data of the road network. The core problem is to capture the spatio-temporal relationship of the traffic flow at each of the monitoring points within the study area. In this paper, we adopt a multi-step prediction strategy of 'multiple inputs-multiple outputs', which allows for the acquisition of more detailed prediction information while widening the prediction time span. In this study, the selection of specific time steps is determined by means of an analysis of the time-variation characteristics of traffic flow in the study area. The definitions that are pertinent to this discussion are presented below.

1) Feature matrix: The traffic flow information at time series is utilized as the feature matrix, denoted as $\boldsymbol{Q}_t \in \mathbb{R}^{N\times 1}$, where N represents the number of monitoring points within the traffic network. The model's input sequence is defined as $\{\boldsymbol{Q}_{t-l+1}, \boldsymbol{Q}_{t-l+2}, \cdots, \boldsymbol{Q}_t\} \in \mathbb{R}^{l\times N\times 1}$, with l indicating the size of the input time window. The output sequence produced by the model is $\{\boldsymbol{Q}_{t+1}, \boldsymbol{Q}_{t+2}, \cdots, \boldsymbol{Q}_{t+k}\} \in \mathbb{R}^{k\times N\times 1}$, which represents the predicted traffic flow for the next k time steps.

2) Spatial relationship matrix: The structure of the road network location map, which depicts the positioning of monitoring points, is represented as $\boldsymbol{G} = (V, E, \boldsymbol{A})$. This is the topological framework of the transportation network. The set $V = \{v_1, v_2, \cdots, v_n\}$ represents the monitoring points, while the set E denotes edges. The adjacency matrix $\boldsymbol{A}$, which is a matrix of dimensions $\mathbb{R}^{n\times n}$, encapsulating the node connectivity relationships. For any two points $v_i, v_j \in V$ and $(v_i, v_j) \in E$, the nodes are connected, and the element a_{ij} in the adjacency matrix is 1, otherwise it is 0.

3) Problem formulation: The concept of the prediction task can be understood as the process of learning a mapping function, designated as f, from input historical traffic data $\{\boldsymbol{Q}_{t-l+1}, \boldsymbol{Q}_{t-l+2}, \cdots, \boldsymbol{Q}_t\}$ and a known spatial relation matrix, designated as $\boldsymbol{A}$. This function is then used to predict future traffic flow in several specific time intervals. As illustrated in Eq.(1):

$$\{\boldsymbol{Q}_{t+1}, \boldsymbol{Q}_{t+2}, \cdots, \boldsymbol{Q}_{t+k}\} = f((\boldsymbol{Q}_{t-l+1}, \boldsymbol{Q}_{t-l+2}, \cdots, \boldsymbol{Q}_t), \boldsymbol{G}) \quad (1)$$

2.2 Architecture of ISTA-Transformer Model

The transformer model was initially developed for the purpose of natural language processing and is comprised of two principal components: the encoder and the decoder. The encoder is composed of the following layers, arranged from top to bottom: The remaining layers are as follows: Multi-Head Self-Attention, Add & Norm (residual connection and layer normalization), Feed-Forward Neural Network, and another Add & Norm layer. The decoder comprises the following layers, arranged from top to bottom: Masked Multi-Head Self-Attention, an Add & Norm layer, a Feed-Forward Neural Network, and another Add & Norm layer.

In this paper, PyTorch is employed to construct an enhanced spatio-temporal attention transformer model, designated the transformer (ISTA-Transformer). The model's comprehensive architectural framework is depicted in Fig. 2.

The complete encoder architecture is comprised of a spatio-temporal embedding layer and K encoders. Each encoder is constituted of multiple encoder layers, which in turn comprise multi-head self-attention mechanisms and feed-forward neural networks. The historical sequence S and the spatial correlation matrix are taken as input in order to obtain a high-order spatio-temporal feature embedding representation $\mathrm{STEI}^{(K)}$. The self-attention mechanism enables the model to consider dependencies both in a longitudinal manner, within the same time sequence, and laterally, across different time sequences from disparate traffic monitoring points. This mechanism

enables the model to learn dynamic patterns within time sequences while also capturing common features or associations across different traffic monitoring points.

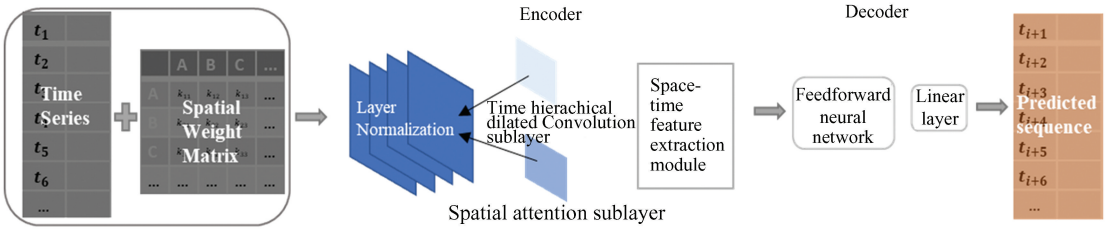


Fig. 2 Architecture of ISTA-Transformer

In the proposed model, the traditional transformer is extended by the incorporation of spatial self-attention sub-layers and temporal hierarchical diffusion convolution sub-layers within each encoder layer. The incorporation of spatial self-attention sub-layers enhances the model's capacity to discern spatial dependencies across disparate monitoring points, enabling each point to attend to other points within the network. This integration enables the model to effectively learn spatial correlations and dependencies, which are crucial for understanding traffic patterns across different locations.

The temporal hierarchical diffusion convolution sub-layers are introduced with the objective of capturing temporal dependencies in a more effective manner, through the application of diffusion convolution across a range of time scales. The hierarchical approach permits the model to extract both short-term variations and long-term trends within the traffic flow data. The combination of these temporal convolution layers with the self-attention mechanism enables the model to better capture complex temporal patterns that traditional self-attention might miss.

The combination of these specialized sub-layers with the traditional transformer architecture allows our model to address the spatial and temporal dimensions of the data simultaneously, resulting in a more comprehensive and accurate spatio-temporal feature extraction.

A residual connection is added after each sub-layer to ensure effective gradient backpropagation, while layer normalization is applied to stabilize training. Finally, the encoder feeds the spatio-temporal feature representation of the historical sequence into each decoder in the lower layers.

2.2.1 Spatio-temporal embedding layer

In this paper, a learnable embedding matrix is employed to encode the spatial information, thereby generating the corresponding spatial embedding, which is then integrated with the traffic flow data. The precise formula is given by Eq.(2):

$$\mathrm{STEI}^{(0)} = \mathrm{Add}(\mathrm{Concat}(\boldsymbol{X}, \boldsymbol{W}_{\mathrm{sp}}), \boldsymbol{S}_t) \quad (2)$$

Where $\boldsymbol{X}$ represents the traffic feature matrix of monitoring point, $\boldsymbol{W}_{\mathrm{sp}}$ is the spatial correlation coefficient matrix, $\boldsymbol{S}_t$ is the learnable temporal embedding, Add is the element-wise addition function, and Concat is the concatenation function that merges traffic feature matrix ($\boldsymbol{X}$) and the spatial correlation coefficient matrix $\boldsymbol{W}_{\mathrm{sp}}$ along a specified dimension.

The spatio-temporal embedding information $\mathrm{STEI}^{(0)}$ is used as the input to the first encoder layer. Residual skip connections are employed to prevent the vanishing embedding features as the number of encoders or decoders increases. The results of the embedding for each node are connected to the node features and then projected into the model dimension.

The combination of $\boldsymbol{X}$ and $\boldsymbol{W}_{\mathrm{sp}}$, followed by the addition of $\boldsymbol{S}_t$, results in the generation of an initial feature representation that encompasses both spatial and temporal data. This constitutes the initial stage of the feature extraction process. Once the preliminary embedding has been produced, the subsequent step is to incorporate the requisite spatial information into the feature matrix, thus enabling the model to utilize spatial dependencies more effectively.

The spatio-temporal embedding layer of the encoder establishes a connection between the flow data from each traffic monitoring point and the spatial correlation coefficient matrix between these points. The spatial correlation coefficient matrix embedding is derived from a preliminary analysis of the spatial relationships between traffic monitoring points and is employed for the computation of spatial attention.

By applying the matrix $\boldsymbol{W}_{\mathrm{sp}}$ to transform the input

feature matrix $\boldsymbol{X}$, we obtain:

$$\boldsymbol{X}' = \boldsymbol{X} \times \boldsymbol{W}_{\mathrm{sp}} \tag{3}$$

This transformed feature matrix $\boldsymbol{X}'$ is subsequently fed into the self-attention layers, providing a representation that incorporates the learned spatial correlations. Furthermore, in convolutional layers designed to capture spatial patterns, the spatial weight matrix can be applied to adjust the convolutional filters, enhancing the spatial feature extraction process. This integration of the spatial weight matrix within different components of the encoder architecture ensures a more comprehensive and accurate capture of both spatial and temporal dependencies in the traffic data.

2.2.2 Attention mechanism

The multi-head attention mechanism represents a pivotal element of the transformer model, devised to address both local and global dependencies inherent to sequential data. This is achieved by the model learning multiple sets of different attention weights (heads), which are then merged to form a composite representation of the input data. This enables the model to concurrently concentrate on disparate elements of the input sequence across distinct subspaces, thereby facilitating more efficacious capture of dependencies within the sequence. This attention mechanism in the ISTA-Transformer, through the learning of spatial dependencies between disparate monitoring points and the selection of pivotal dependencies between time series, enhances the inefficient dependencies between time steps observed in the conventional self-attention mechanism. In the multi-head attention mechanism, the input query (Q), key (K) and value (V) matrices are processed through multiple heads. For the ith header, the computation process is as follows:

$$\mathrm{Attention}^{i}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\,\boldsymbol{W}_i^{Q}\,(\boldsymbol{K}\,\boldsymbol{W}_i^{K})^{\mathrm{T}}}{\sqrt{d_k}}\right)\cdot \boldsymbol{V}\,\boldsymbol{W}_i^{V} \tag{4}$$

where $\boldsymbol{W}_i^{Q}$, $\boldsymbol{W}_i^{K}$, and $\boldsymbol{W}_i^{V}$ are the weight matrices of the query, key, and value of the ith head, and d_k is the dimension of the key vector. In ISTA-Transformer, the multi-head attention mechanism facilitates the capture of both global and local dependencies, while also enhancing spatial self-attention through the introduction of a spatial weight matrix, denoted as sp. In particular: For each subject, the calculation yields a standardized attention score, designated as $\boldsymbol{A}^{i}$:

$$\boldsymbol{A}^{i} = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\,\boldsymbol{W}_i^{Q}\,(\boldsymbol{K}\,\boldsymbol{W}_i^{K})^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{5}$$

The attention matrix is adjusted using the spatial weight matrix $\boldsymbol{W}_{\mathrm{sp}}$:

$$\boldsymbol{A}'^{i} = \boldsymbol{A}^{i} \odot \boldsymbol{W}_{\mathrm{sp}} \tag{6}$$

where $\odot$ denotes element-wise multiplication. This modification ensures that the attention scores are influenced by the actual spatial dependencies, allowing the model to emphasize more relevant spatial relationships in the attention process. The normalized attention scores $\boldsymbol{A}'^{i}$ are then used to compute the weighted sum of the value vectors, facilitating a more contextually aware representation.

Apply the adjusted attention matrix to the value vectors and merge the outputs of all the heads:

$$\begin{aligned}&\mathrm{MultiHeadAttention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \\ &\quad \mathrm{Concat}(\mathrm{Attention}^{1}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}),\mathrm{Attention}^{2}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}),\cdots,\\ &\quad \mathrm{Attention}^{n}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}))\,\boldsymbol{W}^{0}\end{aligned} \tag{7}$$

where n is the number of heads and $\boldsymbol{W}^{0}$ is the weight matrix used to combine the outputs of all the heads.

The multi-head attention mechanism, a core component of the ISTA-Transformer, markedly enhances the model's capacity to discern intricate spatio-temporal patterns. This is achieved by enabling the model to simultaneously learn diverse attention patterns and spatial dependencies. The incorporation of the spatial weight matrix $\boldsymbol{W}_{\mathrm{sp}}$ enables each attention head to adjust the attention scores in accordance with the spatial correlations, thereby enhancing the model's capacity for spatio-temporal feature extraction. This enhancement not only optimizes the attention mechanism but also improves the model's capacity to process real monitoring point data.

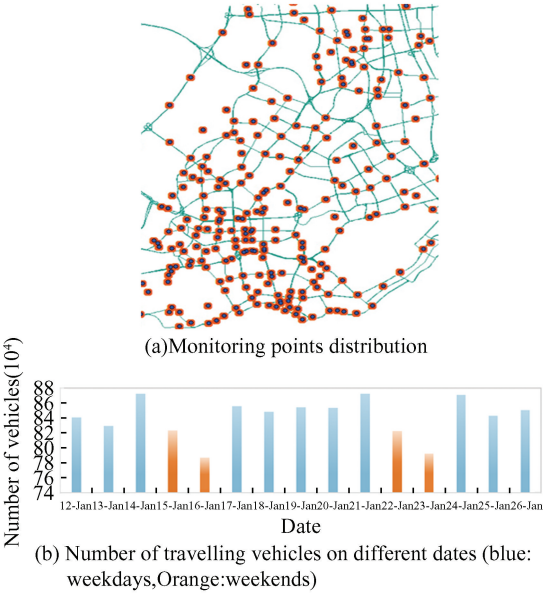
3 Case Study

3.1 Dataset Processing

This paper presents an analysis of the road network in the Shibei District and Shinan District of Qing dao, China. Fig. 3 (a) illustrates that the distribution of traffic monitoring points and monitoring devices on the road network is relatively dense, encompassing the primary traffic nodes within the designated research area. Following an initial screening process, whereby data from certain monitoring points was excluded due to errors resulting from device ageing or damage, this study selected a total of 620 monitoring points for the analysis of 15

days of traffic data (from 12 January to 26 January, 2022). The data set comprises approximately one million vehicle travel records for each day. The original data set comprises a series of records, including the location, ID number, licence plate number and monitoring time of each monitoring point. These records are transformed into traffic flow data, representing the number of vehicles passing through a specific direction of a road within a given time period.

Traffic flow is defined as the number of vehicles passing through a road monitoring point during a specific time interval. In order to gain a preliminary understanding of the temporal characteristics of traffic flow in Qingdao, this study counted the daily number of vehicles travelling on weekdays and weekends during the study period, as shown in Fig. 3(b). As can be seen from the figure, there is a significant difference between the number of vehicles travelling on weekdays and weekends, and it can be deduced that the overall flow on weekdays is higher than that on weekends. Therefore, in order to ensure that the training model can learn a stable spatio-temporal pattern, it is necessary that the training set fully covers the characteristics of traffic flow changes in different periods, such as the traffic difference between weekdays and weekends. In addition, the test set should select data in subsequent periods in order to test the generalization ability of the model on previously unseen data and evaluate its prediction accuracy for new traffic flow conditions.



(a)Monitoring points distribution

(b) Number of travelling vehicles on different dates (blue: weekdays,Orange:weekends)

Fig. 3 Traffic flow distribution feature

Furthermore, it should also include both weekdays and weekends. Consequently, the data from 12 January 2022 (Wednesday) to 21 January 2022 (Friday) are allocated to the training sets, and the data from 22 January 2022 (Saturday) to 26 January 2022 (Wednesday) are allocated to the test sets. This partitioning method is a common approach in the field of traffic flow prediction, and it has been demonstrated to be an effective means of evaluating model performance.

Of the aforementioned traffic monitoring points, Xianggang Middle Road and Fuzhou South Road represent two of the most significant traffic arteries in the Shinan District. The intersection is situated at the core of the densely interconnected road network in the Shinan District. The location is in close proximity to two major commercial districts and a multitude of office buildings. The mean daily traffic volume at this intersection is approximately 120000 vehicles, with an average daily congestion time of 1.5 h. Moreover, the surrounding intersections are closely related to this one. This is a relatively typical urban area road traffic monitoring point.

The data from the traffic monitoring points for the Fuzhou South Road - Xianggang Middle Road intersection (ID 601018114050) is employed as the prediction target. A spatial autocorrelation analysis is conducted on the traffic flow within a 1.2 km radius centered around the target point to identify groups of monitoring points with strong correlations, as illustrated in Fig. 4. The model is trained to complete traffic flow prediction for this specific traffic monitoring point.

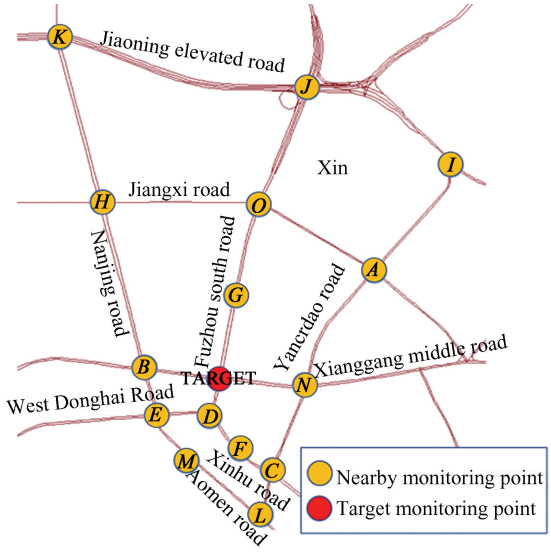


Fig. 4 Monitoring points group

3.2 Traffic Spatio-temporal Feature Analysis

1) Time features: Traffic flow can be classified into distinct intervals based on the temporal resolution of the data, including daily, hourly, and 15 min traffic flow. Fig. 5 depicts the fluctuations in traffic volume at identical traffic monitoring points with varying sampling intervals. Fig. 5(a) depicts traffic statistics for a 5-minute interval. It can be observed that a very short sampling time window leads to significant fluctuations in traffic volume with relatively low absolute values, which is not conducive to prediction and practical application. Conversely, an overly long sampling time window can result in a reduction in the number of data samples. Therefore, a 15 min sampling interval, as illustrated in Fig. 5(c), can effectively smooth the traffic monitoring points traffic fluctuation curve while preserving relatively rich detail features.

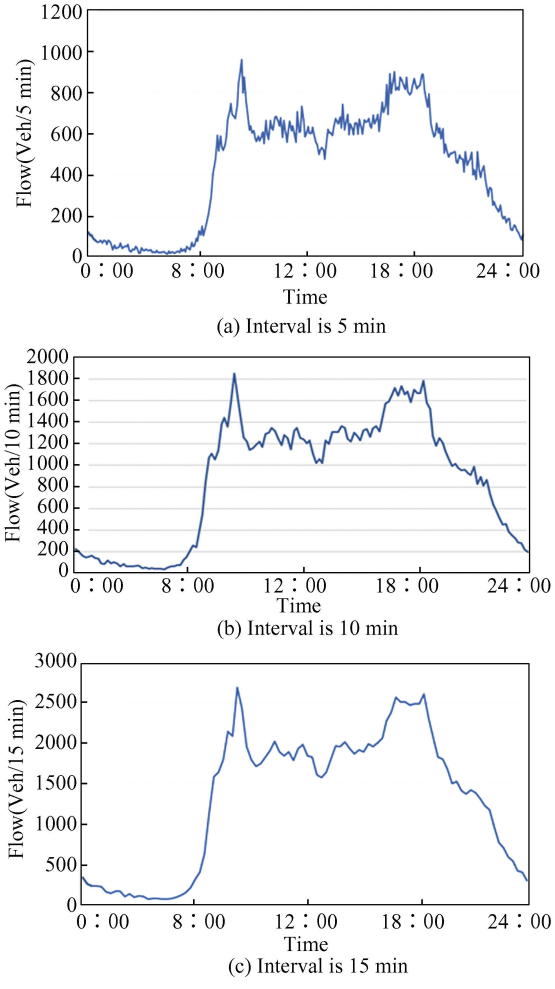


Fig. 5 Traffic volume at different intervals

The individual elements of the input datasets are the traffic flows at the 16 monitoring points, recorded at 15 min intervals. Fig. 6 illustrates the traffic flow in three-dimensional format for 12 January 2022 from 5∶00 to 22∶00 at the choke point for a 15-analysis interval. It is evident that there are discrepancies in the traffic flow patterns observed at different monitoring points within the same time frame. Additionally, the traffic flow at a given monitoring point exhibits considerable fluctuations over time. Accordingly, this paper investigate the spatial and temporal relationship between the surrounding chokepoint traffic and the predicted target chokepoint traffic, with the objective of making a more accurate prediction of the target chokepoint traffic.

Fig. 6 illustrates the historical flow data for the 16 selected monitoring points at 15-minute statistical intervals, and it can be seen that there are significant differences in the flow values between the points, but they are roughly similar in terms of periodicity, making it necessary to explore the effect of coupling between the flow and spatial location of the surrounding and target monitoring points. The following section will further examine the spatial correlation between the traffic flow at different monitoring points and the flow at the target prediction points.

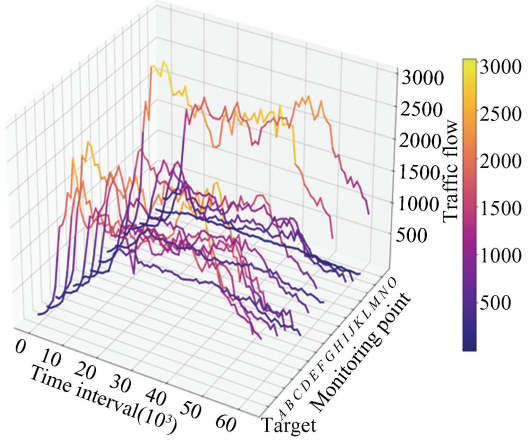


Fig. 6 Traffic flows at group of monitoring points (interval is 15 min)

2) Spatial features: The configuration of urban road traffic networks is inherently complex, characterized by the intersecting of road segments and the relatively short distances between roads. The traffic flow on a given road segment can be affected by the traffic flow on the upstream segment, which in turn can affect the downstream traffic flow. The congestion or smooth flow of traffic on the downstream segment is likely to be influenced by the traffic conditions on the upstream segment. This

illustrates that urban road traffic flow displays spatial correlation. As illustrated in Fig. 6, the flow change curves of monitoring points C, D, E, and F, situated on the same main road, exhibit a notable degree of resemblance.

In this study, the Pearson coefficient is employed to quantify the relationships between the target traffic monitoring points and their surrounding traffic monitoring points. This coefficient incorporates a spatial weight matrix, $\boldsymbol{w}_{ij}$, to derive spatial correlation results. The formula for calculation is presented in Eq.(8):

$$\rho_{ij} = \boldsymbol{w}_{ij} \frac{E[(X_i(t) - \bar{X}_i)(X_j(t) - \bar{X}_j)]}{\sqrt{\sigma_{X_i}\sigma_{X_j}}} \tag{8}$$

where X_i, X_j are the time series traffic flow data for traffic monitoring points i, j, respectively. $X_i(t)$ represents the time series traffic flow data at traffic monitoring points i at time t. $X_j(t)$ represents the time series traffic flow data at traffic monitoring points j at time t. E denotes the mathematical expectation. σ represents the standard deviation.

Fig. 7 depicts the spatial correlation coefficient diagram between the research objects. This diagram illustrates the spatial correlations between the various traffic monitoring points under study, based on the spatial correlation analysis conducted using the provided formula. The coefficients offer insights into the relationship between the traffic flow at one traffic monitoring point and the traffic flow at neighbouring traffic monitoring points, thereby facilitating an understanding of the spatial relationships within the road network.

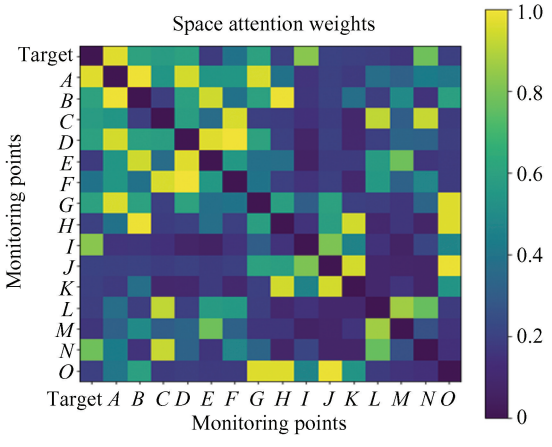


Fig. 7 Spatial correlation coefficient

A sliding window is a data processing technique employed for the generation of sample sequences, which are subsequently utilized for the analysis of time series data. In this article, the sliding window method is employed to transform the original time series data into a set of data points suitable for training and testing deep learning models. This involves the creation of a time window with a size of 12 time steps, wherein the observations within this window serve as input features, and the traffic values for the subsequent four time steps are used as the target sequence (see Fig. 8). Following the application of the sliding time window for the division of the samples, the dimensions of the samples processed by the algorithm employed in this study are as follows: (number of samples, length of time window, number of monitoring points). In other words, each sample comprises data from 12 time steps with 16 features per time step, while the dimensions of the set labels are (number of samples, number of predicted time steps, number of predicted points). Consequently, each label contains data from 4 time steps with 1 feature per time step.

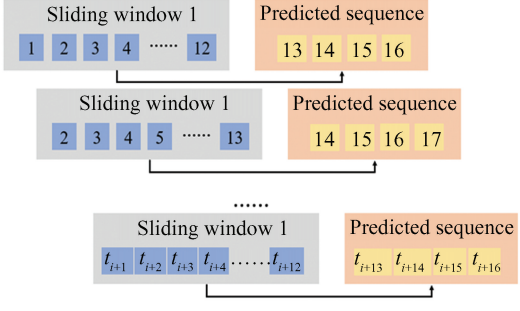


Fig. 8 Slide prediction time window

3.3 Model Evaluation

To evaluate the model's performance, error functions are commonly used as evaluation metrics for prediction results, including Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2).

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\% \tag{9}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{10}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{11}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2} \tag{12}$$

To evaluate the performance of the ISTA-Transformer model, it was compared with several baseline models, including the Transformer, GRU (Gated Recurrent Unit), CNN (Convolutional Neural Network), Long short-term Memory Network (LSTM), Sequential Neural Network (Seq2Seq) and LightGBM.

1) Transformer: Configured with 6 layers for both the encoder and decoder, 8 attention heads, a hidden dimension of 512, a feedforward dimension of 2048, and a dropout rate of 0.1.

2) Gated Recurrent Unit (GRU): A variant of recurrent neural networks (RNNs) that simplifies the architecture compared with Long short-term Memory (LSTM) models. The GRU model used here has 2 layers, 64 hidden units, and a dropout rate of 0.2.

3) Convolutional Neural Network (CNN): Designed to process and learn from spatial and temporal patterns in data. For this comparison, the CNN was configured with 3 convolutional layers, a kernel size of 3, a stride of 1, and 64 filters.

4) Long short-term Memory Network (LSTM): A type of recurrent neural network designed to capture long-term dependencies in sequential data. The LSTM model used in this evaluation has 2 layers, 64 hidden units per layer, and a dropout rate of 0.2.

5) Sequential Neural Network (Seq2Seq): A neural network architecture designed to address sequence-to-sequence problems. The model comprises two principal components, namely the encoder and decoder. The LSTM layers of the encoder and decoder are configured to have 64 hidden units, and the network has two LSTM layers, a dropout rate of 0.2 is employed. The model was trained for 200 cycles, with each batch comprising 32 samples.

6) LightGBM: An efficient gradient boosting framework for handling large-scale datasets and high-dimensional features. Number of estimators is 200, learning rate is 0.01, number of leaves is 31, random_state is 42, stopping_rounds is 50. Additionally, a logging callback function is utilized to monitor the evaluation metrics throughout the training process and log the model performance every 10 rounds.

These baseline models were used to assess the performance of the ISTA-Transformer model by providing a range of different forecasting approaches and capabilities.

4 Results and Discussion

4.1 Model Training

As illustrated in Fig. 9, the mean-square error loss function curve is depicted when the ISTA-Transformer model processes normalized traffic data. The raw loss values are displayed with a blue curve, showing all fluctuation details, while the red curve represents the smoothed loss trend. In the figure, a 5-epoch moving window is used to calculate the mean and standard deviation, with the red shaded area indicating the fluctuation range of ±1 standard deviation. It can be observed that the loss function exhibited significant growth during the initial 10 epochs, followed by slight fluctuations within the subsequent 100 epochs. Subsequently, it rapidly converged. This suggests that the model is a suitable tool for traffic flow prediction tasks.

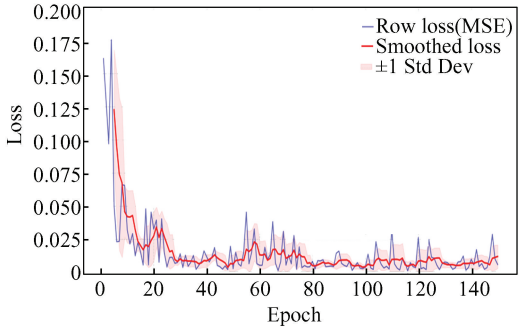


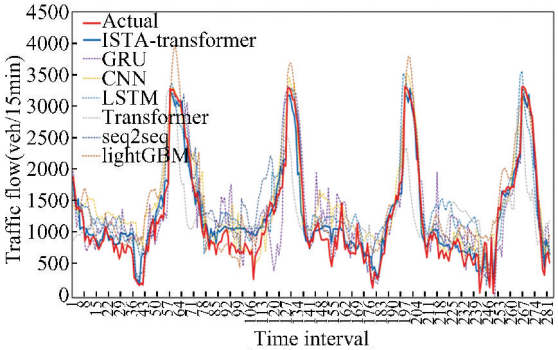
Fig.9 Loss curve of the ISTA-Transformer model

4.2 Prediction Results

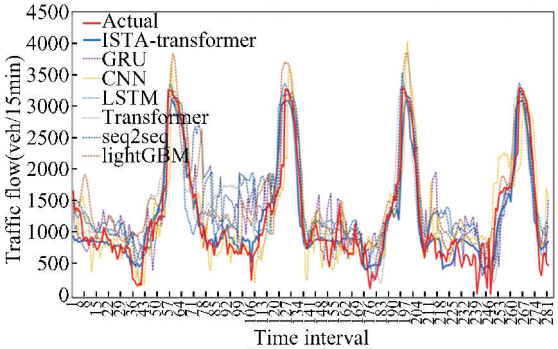
The ISTA-Transformer model was evaluated in comparison with all baseline models for performance assessment on the monitoring points traffic dataset, including four sequential time steps (15, 30, 45, and 60 min), which are four consecutive time intervals in 15-minute steps, collectively referred to as "four-step prediction". This formulation enables the model to consider dependencies across multiple future intervals, thereby improving forecasting stability over different time horizons. The results are presented in Fig. 10.

As demonstrated in Fig. 10, the accuracy of the predictions varies at different time steps, with all models demonstrating superior prediction at the first time step in comparison to the subsequent time steps. This finding suggests that, with limited historical data, each model exhibits greater capacity to predict traffic flow in the initial 15 min of the future. However, as the time step increases, the discrepancy

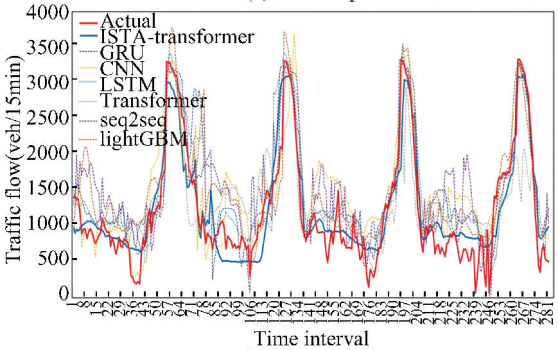
between the predicted values and the true values output by each model becomes significantly larger, indicating a decline in accuracy over longer time spans. This may be attributable to the elevated uncertainty surrounding traffic conditions, in conjunction with the time-dependent and error-passing effects inherent in the model, which render forecasting more arduous.



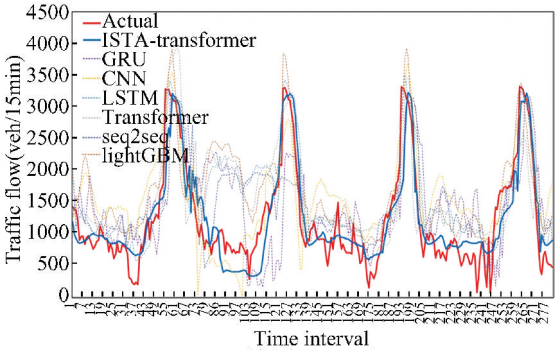
(a) Time step 1



(b) Time step 2



(c) Time step 3



(d) Time step 4

Fig.10 Actual values versus predicted values

Furthermore, the majority of baseline models exhibit an overly optimistic prediction trend during the off-peak hours, and for peak periods, baseline models such as LightGBM consistently show predictions that are significantly higher than the true values. The baseline model also exhibits an outlier at the fourth time step, deviating significantly from the actual prediction and accompanied by overall anomalous fluctuations (the prediction is close to 0). In contrast, the ISTA-Transformer model demonstrates a superior capacity in capturing traffic flow trends, exhibiting substantial consistency between predicted and actual values during periods of significant traffic flow variability. The predicted and actual values are highly proximate to each other, both during peak and off-peak traffic flow periods.

In order to provide a clearer indication of the error level of the predicted value, the inverse normalization process was carried out to convert the normalized data back to the original scale. The four evaluation indicators are then calculated using Eqs. (9)-(12) (see Table 1 and Fig.11). Obviously, for each time step, the ISTA-Transformer model consistently outperforms the other baseline models on all four evaluation metrics. A review of the overall performance across the different metrics shows that the ISTA-Transformer model gives the most favourable predictions in the early time steps.

When predicting traffic flow in the first 15 min, the ISTA-Transformer model has a MAPE value of 0.22, which is significantly lower than the other models. This shows that it is better at capturing the dynamic characteristics of traffic flow in the short term. For traffic flow predictions beyond the initial 30-minute interval, the ISTA-Transformer model continues to have the most favourable MAPE value of 0.29. For a forecast time step of 45 min, the ISTA-Transformer model has a MAPE of 0.37, which is slightly higher, but still shows superior performance compared with the other models. This result shows that ISTA-Transformer is still effective in capturing the spatio-temporal dependence of the data as time steps increase, while other models begin to show larger errors. When the longest time step was 60 min, the MAPE of all models increased significantly. The ISTA-Transformer model has a MAPE of 0.38, which is the lowest of all the models. This shows that even when making predictions over a longer time horizon, the ISTA-Transformer model still shows superior robustness and generalization. This shows that ISTA-Transformer can effectively adapt to short time

intervals and traffic data with significant periodicity. The model consistently achieves higher R^2 values at different time steps, demonstrating its strong ability to adapt to fluctuations in traffic data and identify potential patterns, confirming its advantages in traffic flow prediction tasks.

Table 1 Prediction error results for different time steps for each model

Time step	Model	MAPE	MSE	RMSE	R^2
Time step 1 (T=15 min)	ISTA-Transformer	0.22	40096	200.2	0.94
	GRU	0.40	136624	369.6	0.79
	CNN	0.46	136579	369.6	0.79
	LSTM	0.48	154084	392.5	0.76
	Transformer	0.49	335850	579.5	0.48
	Seq2seq	0.24	75849	275.4	0.88
	LightGBM	0.48	124797	353.3	0.81
Time step 2 (T=30 min)	ISTA-Transformer	0.29	78715	280.6	0.88
	GRU	0.53	178127	422.1	0.72
	CNN	0.42	172824	415.7	0.73
	LSTM	0.54	228748	478.3	0.64
	Transformer	0.48	192789	439.1	0.70
	Seq2seq	0.53	231935	481.6	0.64
	LightGBM	0.56	203170	450.7	0.68
Time step 3 (T=45 min)	ISTA-Transformer	0.37	96521	310.7	0.85
	GRU	0.59	257160	507.1	0.60
	CNN	0.59	230471	480.1	0.64
	LSTM	0.48	186785	432.2	0.71
	Transformer	0.56	234787	484.5	0.64
	Seq2seq	0.55	217591	466.5	0.66
	LightGBM	0.58	212812	461.3	0.67
Time step 4 (T=60 min)	ISTA-Transformer	0.38	120570	347.2	0.81
	GRU	0.77	536556	732.5	0.17
	CNN	0.79	350758	592.2	0.46
	LSTM	0.62	312047	558.6	0.52
	Transformer	0.78	409517	639.9	0.37
	Seq2seq	0.69	315262	561.5	0.51
	LightGBM	0.75	392833	626.8	0.39

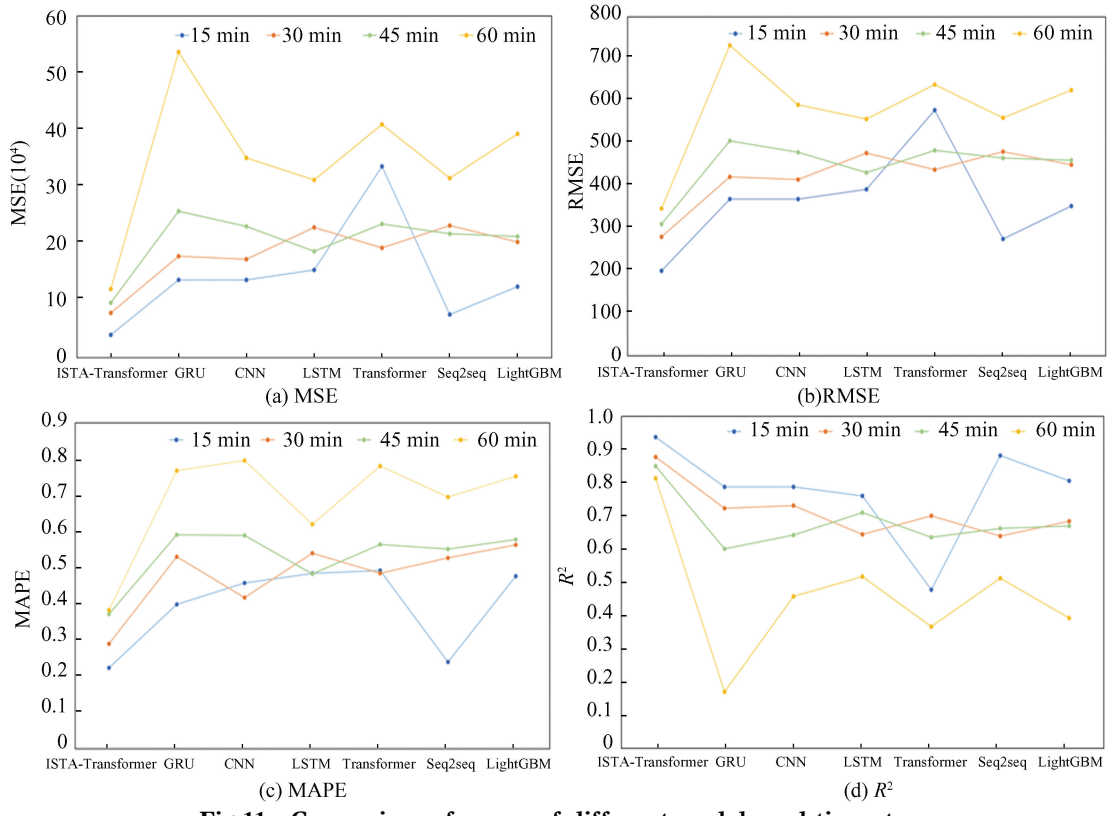


Fig.11 Comparison of errors of different models and time steps

5 Conclusions and Future Work

The ISTA-Transformer traffic flow prediction model, as proposed in this paper, has been designed with the objective of facilitating effective multi-step prediction of short-term traffic flow. This model incorporates the spatial attention weight matrix into the traffic flow prediction process, thereby addressing the spatial dimension of traffic data and simulating the highly nonlinear spatial correlation between road segments. On this basis, a spatio-temporal attention mechanism is introduced, whereby the spatio-temporal features extracted by the encoder are superimposed. The historical traffic flow data of 16 monitoring points within the selected research range is used as the feature input, with a time step of 15 min and a sliding time window to divide training samples. This enables the prediction of the traffic flow value of the target monitoring point in the next four time steps. The experimental results on the Qingdao traffic monitoring dataset demonstrate that the ISTA-Transformer model proposed in this paper outperforms the six selected baseline models. In the context of multi-step short-term prediction, the models under consideration are the GRU, CNN, LSTM, Transformer, Seq2Seq and LightGBM. The ISTA-Transformer model demonstrated superior performance compared with the other baseline models in terms of MSE, RMSE, MAPE, and R^2 error evaluation indicators. In the case study, the ISTA-Transformer model, trained on a limited amount of historical data, demonstrated robust predictive capabilities in the test set. It is evident that the model's performance is most precise at shorter time intervals (15 – 30 min), indicating that the model has the potential for application in the field of traffic flow prediction. Furthermore, it can provide high accuracy while also conserving computing resources. While an increase in prediction errors is observed for longer horizons (45 – 60 min), which indicates that while ISTA-Transformer effectively captures spatio-temporal dependencies, the incorporation of advanced techniques such as adaptive attention mechanisms or hybrid models might be necessary to enhance the accuracy of long-horizon forecasting.

In the future, this research intends to consider additional factors, such as weather and traffic accidents, in traffic predictions and translate these influencing factors into learnable features. Additionally, further analysis will be conducted on the distinct characteristics of different time steps in multi-step prediction. Exploration of techniques such as dynamic time-step adjustment and uncertainty quantification will be undertaken to enhance the robustness of long-horizon forecasts. Moreover, the objective is to enhance fine-grained forecasting of peak hour traffic, bringing the model closer to real traffic scenarios, thereby improving its performance in multivariate forecasting.

References

[1] Yuan H, Li G. A survey of traffic prediction: from spatio-temporal data to intelligent transportation. Data Science and Engineering, 2021,6(1): 63-85. DOI:10.1007/s41019-020-00151-z.

[2] Wang X, Sun F, Ma X, et al. Short-term traffic flow prediction based on vehicle trip chain features. Transportation Letters, 2024, 17(1): 157-168. DOI: 10.1080/19427867.2024.2334100.

[3] Ahmed M S, Cook A R. Analysis of freeway traffic time-series data by using box-jenkins techniques. Transportation Research Record, 1979,722:1-9.

[4] van der Voort M, Dougherty M, Watson S. Combining Kohonen maps with ARIMA time series models to forecast traffic flow. Transportation Research Part C: Emerging Technologies,1996,4(5):307-318. DOI:10.1016/S0968-090X(97)82903-8.

[5] Lee S, Fambro D B. Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting. Transportation Research Record, 1999,1678:179-188. https://doi.org/10.3141/1678-22.

[6] Williams B M. Multivariate vehicular traffic flow prediction: Evaluation of ARIMAX modeling. Transportation Research Record: Journal of the Transportation Research Board, 2001,1776(1): 194-200. DOI:10.3141/1776-25.

[7] Kamarianakis Y, Prastacos P. Forecasting traffic flow conditions in an urban network: Comparison of multivariate and univariate approaches. Transportation Research Record, 2003,1857:74-84. DOI:10.3141/1857-09.

[8] Williams B M, Hoel L A. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. Journal of Transportation Engineering, 2003, 129(6): 664-672. DOI:10.1061/(ASCE)0733-947X(2003)129:6(664).

[9] Karami Z, Kashef R. Smart transportation planning: Data, models, and algorithms. Transportation Engineering, 2020, 2:100013. DOI:10.1016/j.treng.2020.100013.

[10] Luan X, Yang B, Zhang Y. Structural hierarchy analysis of streets based on complex network theory. Geomatics and Information Science of Wuhan University, 2012, 37(6): 728-732.

[11] Medina-Salgado B, Sánchez--DelaCruz E, Pozos-Parra P, et al. Urban traffic flow prediction techniques: A review. Sustainable Computing: Informatics and Systems, 2022, 35:100739. DOI:10.1016/j.suscom.2022.100739.

[12] Lv Y, Duan Y, Kang W, et al. Traffic flow prediction with big data: A deep learning approach. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(2): 865-873. DOI:10.1109/TITS.2014.2345663.

[13] Xia D, Yang N, Jian S, et al. SW-BiLSTM: a Spark-based weighted BiLSTM model for traffic flow forecasting. Multimedia Tools and Applications, 2022, 81(17): 23589-23614. DOI:10.1007/s11042-022-12039-3.

[14] Chauhan N S, Kumar N, Eskandarian A. A novel confined attention mechanism driven Bi-GRU model for traffic flow prediction. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(8): 9181-9191. DOI: 10.1109/TITS.2024.3375890.

[15] Luo Y, Zheng J, Wang X, et al. GT-LSTM: A spatio-temporal ensemble network for traffic flow prediction. Neural Networks, 2024, 171: 251-262. DOI: 10.1016/j.neunet.2023.12.016.

[16] Zhang Z, Li M, Lin X, et al. Multistep speed prediction on traffic networks: A deep learning approach considering spatio-temporal dependencies. Transportation Research Part C: Emerging Technologies, 2019, 105: 297-322. DOI: 10.1016/j.trc.2019.05.039.

[17] Ali A, Zhu Y, Zakarya M. Exploiting dynamic spatio-temporal graph convolutional neural networks for citywide traffic flows prediction. Neural Networks, 2022, 145: 233-247. DOI:10.1016/j.neunet.2021.10.021.

[18] He R, Xiao Y, Lu X, et al. ST-3DGMR: Spatio-temporal 3D grouped multiscale ResNet network for region-based urban traffic flow prediction. Information Sciences, 2023, 624: 68-93. DOI:10.1016/j.ins.2022.12.066.

[19] Ma Y, Lou H, Yan M, et al. Spatio-temporal fusion graph convolutional network for traffic flow forecasting. Information Fusion, 2024, 104: 102196. DOI: 10.1016/j.inffus.2023.102196.

[20] He R, Liu Y, Xiao Y, et al. Deep spatio-temporal 3D densenet with multiscale ConvLSTM-Resnet network for citywide traffic flow forecasting. Knowledge-Based Systems, 2022, 250: 109054. DOI: 10.1016/j.knosys.2022.109054.

[21] Yang J, Sun X, Wang R G, et al. PTPGC: Pedestrian trajectory prediction by graph attention network with ConvLSTM. Robotics and Autonomous Systems, 2022, 148: 103931. DOI:10.1016/j.robot.2021.103931.

[22] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Proceedings of the 31th Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010.

[23] Xiao L, Chen H. Spatio-temporal transformer graph network for traffic flow forecasting. Proceedings of the 2024 3rd International Joint Conference on Information and Communication Engineering. Fuzhou: JCICE, 2024: 224-228. DOI:10.1109/JCICE61382.2024.00053.

[24] Hu H-X, Hu Q, Tan G, et al. A multi-layer model based on transformer and deep learning for traffic flow prediction. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(1): 443-451. DOI: 10.1109/TITS.2023.3311397.

[25] Zhan X, Zhang S, Szeto W Y, et al. Multi-step-ahead traffic speed forecasting using multi-output gradient boosting regression tree. Journal of Intelligent Transportation Systems: Technology, Planning, and Operations, 2020, 24(2): 125-141. DOI: 10.1080/15472450.2019.1582950.

[26] Bai L, Yao L, Wang X, et al. Deep spatial-temporal sequence modeling for multi-step passenger demand prediction. Future Generation Computer Systems, 2021, 121: 25-34. DOI:10.1016/j.future.2021.03.003.

[27] Zhao K, Guo D, Sun M, et al. Short-term traffic flow prediction based on hybrid decomposition optimization and deep extreme learning machine. Physica A: Statistical Mechanics and Its Applications, 2024, 647: 129870. DOI: 10.1016/j.physa.2024.129870.