Research Article

An Effective Self-Attention-Based Hybrid Model for Short-Term Traffic Flow Prediction

Zhihong Li ⓘ, Xiaoyu Wang, and Kairan Yang

[1]*Department of Transportation, Beijing University of Civil Engineering and Architecture, Beijing 100044, China*

Correspondence should be addressed to Zhihong Li; lizhihong@bucea.edu.cn

Vehicle exhaust is one of the main sources of carbon emissions. The short-term traffic flow prediction plays an important role in alleviating traffic congestion, optimizing the travel structure, and reducing traffic carbon emissions. The current advanced models of short-term traffic flow prediction are evaluated in this work, especially their inadequacies. To improve the prediction accuracy and ensure fine traffic management, an effective self-attention-based hybrid model is proposed to predict the short-term traffic flow. The proposed model includes an encoder-decoder neural network module and a self-attention mechanism module. The self-attention mechanism module is applied as a feature extraction unit in this hybrid model to enhance the ability of key information capture and to settle the problem on key information disappearing due to the increasing sequence length in traditional models. The dataset of the Guangdong freeway toll station is used for the experimental testing. Compared with several baseline models, the proposed model is more suitable for real-time prediction and can provide highly accurate results. Also, a better interpretability is presented in this proposed model. The experimental results showed that MAE, RMSE, and MAPE of the proposed model are 3.01, 4.38, and 12.99%, respectively. Our new hybrid model gives a higher accuracy than the support vector regression (SVR) model, LSTM neural network-attention (LSTM-attention) model, and temporal convolutional network (TCN) model. It shows that the proposed model in this work is favorable to the short-term traffic flow prediction.

1. Introduction

With the acceleration of urbanization, the number of motor vehicles increases rapidly, and the problem of traffic congestion is more and more serious. The most significant ones are the increment in travel time consumption caused by traffic congestion and environmental pollution from the vehicle exhaust. Reducing the traffic pollution is completely consistent with the national strategy of carbon neutralization and carbon peaking.

In order to optimize traffic distribution and improve traffic efficiency, it is necessary to achieve high accuracy traffic prediction. The current research studies mainly focus on accurate prediction for traffic flow state, or flexible and efficient adjustment for control strategy. The accurate short-term traffic flow prediction can help decision-makers to better understand the traffic flow state, and thus to formulate a more reasonable control strategy. Meanwhile, it can also help drivers to better arrange travel plans and reduce carbon emissions due to traffic congestion and frequent braking times.

The spatiotemporal feature of the traffic flow can be analyzed and predicted by the deep neural network. The accurate traffic flow prediction plays an important role in alleviating the traffic problems brought by the actual development, improving the efficiency of the traffic system and the ability of active prevention and control.

The short-term traffic flow prediction is the prediction of the traffic flow with a time step of fewer than 15 minutes [1]. The short-term traffic flow is characterized by time

series. The traditional time series prediction is a linear fitting method represented by autoregressive integrated moving average (ARIMA) [2, 3]. This method has a problem of poor ability to fit nonlinear data [4]. In addition, there are many research studies on shallow machine learning methods represented by support vector regression (SVR) [5, 6]. However, SVR is slow in processing high-dimensional data, and the choice of different kernel functions will have a great impact on the results. With the rapid development of deep learning algorithms, the recurrent neural network (RNN) [7] and its variants of long short-term memory (LSTM) [8–11] and convolutional neural network (CNN) are also applied to time series prediction and have good results.

However, these models still have some drawbacks: (1) gradient explosion and gradient disappearance for long period time series; (2) with the increase of the sequence length, it is easy to lose information due to the weak extraction ability on key information.

Thus, many scholars put forward a hybrid model that adds an attention mechanism to the model of the CNN or RNN class [12–16], which can effectively mitigate above problems by the ability on extraction critical information [17]. Therefore, a potential self-attention-based hybrid model is proposed in this paper. This model adopts a self-attention mechanism to suppress the loss of long-term time series information and effectively improves the prediction accuracy. The results show that the model performs well based on good validation by the dataset of a freeway toll station in Guangdong.

The major contributions of this paper are summarized as follows:

(1) An effective self-attention-based hybrid model for the short-term traffic flow prediction is proposed. This proposed model is conducive to giving highly accurate prediction results using modest training data from the freeway toll station.

(2) The proposed hybrid prediction method is capable of extracting and learning dominant spatiotemporal features and short-term variations of the traffic flow by the encoder-decoder framework

(3) The effectiveness and efficiency of the proposed model are demonstrated according to one real-worldshort-term traffic prediction case studies

The rest of this paper is given as follows: Section 1 is related to work review, which introduces the time series prediction model in traffic field and its advantages and disadvantages. Section 2 is the basic principle of the model, i.e., the principle of multihead self-attention mechanism and the structure of the model proposed in this paper. Section 3 is the evaluation criteria of data sources and forecast results. Section 4 is the experiment, mainly about the experimental hardware equipment and experimental details. Section 5 is the experimental results and discussion compared with other models. Section 6 is the conclusion.

2. Literature Review

In recent years, scholars have carried out extensive application and research on time series models in the field of transportation, which can be roughly divided into three categories.

The first one is the traditional linear time series prediction model, including ARIMA and the improved models. Most of these improved models focus on enhancing the ability of ARIMA to handle complex data. Shahriari et al. proposed a random dataset generation algorithm considering the influence of randomness and combined it with ARIMA model to effectively improve the model's accuracy [18]. Van Der Voort et al. integrated Kohonen self-organizing graph with ARIMA model [19]. After the Kohonen's initial classification, the nonlinear data processing ability of the model was significantly enhanced. The prediction accuracy raised remarkably. In consideration of the dynamic and nonlinear characteristics of traffic flow changes, Shen et al. combined empirical mode decomposition together with ARIMA model and enhanced its ability to process fluctuating data [20]. Chao et al. proposed a time series analysis method based on ARIMA model structure, which could meet the requirements of the traffic flow dynamic prediction to a certain extent and polish the prediction accuracy base on the problems of too much training data and reducing the value of forgetting factor [21]. In addition, some scholars improved the ARIMA model considering the seasonal [22] and spatial characteristics of the traffic flow [23] and the heteroscedasticity of traffic flow time series [24].

The second type is the shallow machine learning model. Cheng et al. gave a fusion multisource SVR model, which used the maximum Lyapunov index and Bayesian theory to fuse various traffic flow features and then used SVR on regression prediction and obtained good results [25]. Hong et al. combined genetic algorithms-simulated annealing algorithm with SVR to alleviate the problem that was easy to fall into local optimal solution in training [6]. Luo et al.suggested an SVR model modified with particle swarm optimization algorithm and genetic algorithm and calculated parameters through the least square method for optimization of the training speed and model accuracy [26]. Feng et al. proposed an adaptive multikernel SVR short-term traffic flow prediction model in consideration of spatiotemporal correlation and proved that the model could better deal with the complicated traffic flow data [27]. Wei and Liu studied an adaptive SVR model integrated with heuristic algorithm, which greatly improved the computational efficiency [28].

The third category is the deep learning model, including RNN (GRU, LSTM) and CNN models. Deep learning models have shown promising performance in many research areas due to their abilities on complex model nonlinear relationships [29]. Different types of models for the short-term traffic prediction have been developed based on their abilities to capture the spatiotemporal correlations of traffic flow.

One of the commonly used deep learning models is the CNN-based model. The CNN model is often applied in the traffic prediction due to its ability to capture the local dependencies of traffic data and is less sensitive to noise. Chen et al. set up a causal convolution neural network model based on empty convolution [30]. Experimental results showed that the model could effectively learn seasonal and holiday effects in time series. Zhao et al. built a time series classification model based on CNN considering its automatically generation and extraction feature and solved the problem of feature selection dependence in traditional classification methods [31]. In addition, CNNs cannot capture the interdependence between road connections for a specific complex road network structure [8].

Another commonly used deep learning model is the RNN-based model. RNN has been widely used in time series tasks. The LSTM model enlarges the memory of RNN and is suitable to learn from time series data. Cui et al. proposed a deep SBU-LSTM model to predict network-wide traffic speed, which considered forward and backward dependencies in time series data [32]. Cai et al. designed an improved loss function LSTM prediction model for traffic flow data under the influence of non-Gaussian noise, which enhanced the robustness of the model [33]. Ma et al. [34] proposed an improved LSTM model based on LSTM and bi-LSTM. This model integrates bi-LSTM into the prediction model. The results show that the LSTM_BILSTM hybrid prediction model has good accuracy and stability in multistep prediction. Ma et al. [8] proposed a hybrid model of CapsNet + NLSTM to solve the defect that traditional models cannot handle complex spatial relationships when predicting traffic conditions. CaptsNet is used to extract the comprehensive spatial characteristics of the road network, and the NLSTM model is used to capture the hierarchical time dependence of traffic data. The experimental results show that the model has good prediction ability in dealing with complex road networks. Wei et al. [35] set up an auto encoder LSTM (AE-LSTM) model considering the spatial relationship of nodes and applied autoencoder to extract the characteristics of upstream and downstream traffic, thus effectively improved the accuracy of the model. Yang et al. added an attention mechanism module to the LSTM model to deal with the problem of gradient disappearance in the traffic flow data of long-term series [36]. Lea et al. proposed a residual structure neural network model TCN based on full convolution to solve the problems of the gradient stability and inability to process information in parallel in RNN [37].

Although CNN-based and LSTM-based models are much powerful to capture the spatiotemporal relations of traffic flow, the limitations were obvious when there were many study objects. In such cases, deep graph neural networks (GNNs) are superior to extracting features from transportation networks. Research studies mainly focus on the graph convolution network [16, 38–40] and spectral convolution network [41]. The graph convolution neural network builds the traffic graph based on the physical network topology and defines a graph convolution neural network to capture spatial features. Then, the LSTM regression neural network is established by traffic graph convolution (TGC) to predict the traffic space-time speed. Otherwise, the methods of spectral graph convolution and localized spectral graph convolution can capture the space-time characteristics on the network and usually have high prediction accuracy.

Based on application analysis, deep learning models often obtain highly accurate predictions. Therefore, they are now the most advanced and popularly used methods in the field of traffic flow prediction. However, these models also have drawbacks that make them less preferable for certain short-term traffic prediction applications. Firstly, a large amount of data are required usually. Secondly, the time of training is time-consuming and the model may take several days or weeks to train. Finally, the interpretability of the result is lacking.

In order to get accurate and efficient short-term traffic predictions, this study proposes an effective self-attention-based hybrid model framework, in which a multiattentional mechanism can be extended to solve long distance-dependent temporal traffic data and extract dominant information from collinear and correlated data. One real-world traffic state prediction case study in Humen Freeway Toll Station is developed to demonstrate the effectiveness and efficiency of this proposed hybrid model on short-term traffic prediction applications. The proposed model is much suitable for real-time prediction applications and can produce accurate results comparable with those baseline models. Moreover, the proposed model has better interpretability ability.

3. Methodology

3.1. Multiself-Attention Machine Module. The attention mechanism was originally used to solve the problem of long-term sequence dependence in machine translation, in which the performance of machine translation decreased significantly with the increase in the sentence length [16]. It has been widely used in the processing of various time series data. In this work, the attention mechanism is used to exact key contents from a large amount of information, focusing on the crucial parts and ignoring unimportant ones.

The self-attention mechanism is a variant of the attention mechanism, which reduces the dependence on external information and is better at capturing the internal relevance of data or features. First, the weight coefficients are calculated according to query (Q) and key (K), and then value (V) is summed according to the weight coefficients as shown in (1).

The attention model can simultaneously compute the attention functions on a set of queries and pack them into a matrix Q. The keys and values are also packed in the matrices K and V. Attention (Q, K, and V) represents the matrix calculation of the output.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \otimes V,$$

$$\text{Softmax}(x) = \frac{e^{x_i}}{\sum_i^n e^{x_i}}, \tag{1}$$

where n is the length of the input sequence. Three sets of trainable parameter matrices W^Q, W^K, and W^V are obtained by weighting the input data sequence X. The Q, K, and V are obtained by multiplication. The d_k is the dimension of K, and $\otimes$ is the element multiplication, a matrix element multiplication by position.

Considering that a single self-attention mechanism cannot capture the importance of input sequence information comprehensively, therefore, the self-attention mechanism usually adopts a multihead design to identify multiple subspace information simultaneously in different locations. The multiattentional mechanism refers to the parallel connection of several self-attentional mechanism modules. A single self-attention mechanism module is outputted and spliced. It is shown in the following:

$$\text{Multihead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_m)W, \tag{2}$$

$$\text{head}_i = \text{Attention}(XW_i^Q, XW_i^K, XW_i^V), \tag{3}$$

where W is the multiheaded attention weight matrix. The weight matrix of the ith attention head Q, K, and V is $W^Q, W^K,$ and W^V. m is the number of attention heads, and Concat function is used to join the output values calculated by each attention head.

3.2. The Model Structure.

The encoder-decoder framework was first applied to natural language tasks [42]. The basic idea is to use a recurrent neural network to read the input sentence and compress the text information into a fixed dimension code. Another recurrent neural network reads the code and converts it into a sentence with the target language. These two recurrent neural networks are called encoder and decoder, respectively.

The structure of the proposed hybrid model is illuminated in Figure 1. The model first encodes the input data and label data and inputs the encoded data into encoder and decoder modules, respectively. The decoder module outputs final results after full connection layer flattening processing.

The encoding module is the definition of rules for encoding sequences. The sequence numbered even is sinusoidal coded, and the sequence numbered oddly is cosine coded. The coding formula is shown in the following equation:

$$\begin{cases} PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{(2i/d_{\text{model}})}}\right), \\ \\ PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{(2i/d_{\text{model}})}}\right), \end{cases} \tag{4}$$
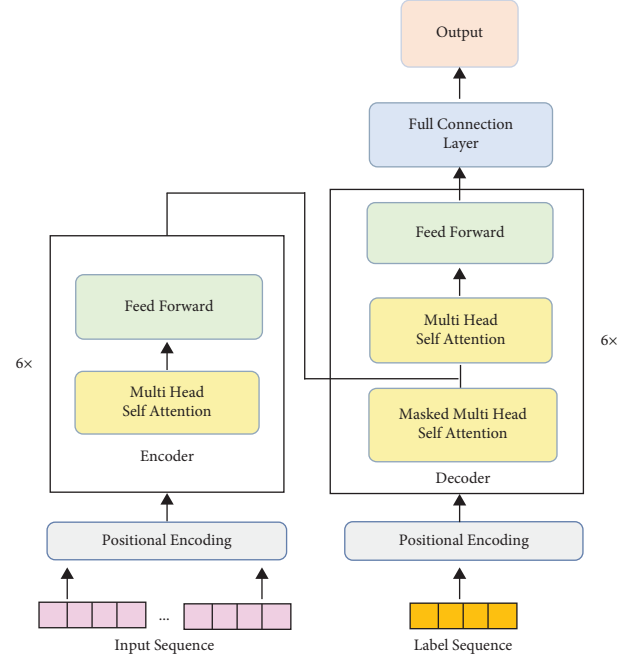

Figure 1: The model structure.

where pos represents the position index of each data in the sequence, i is the ith dimension corresponding to this position, and d_{model} is the dimension of the hidden layer. This encoding format provides great convenience for the model to capture the relative position relationships between words. The application in time series prediction can also play a good effect.

The encoder and decoder of the model are composed of six identical layers. Each layer of the encoder contains two sublayers of multihead self-attention mechanism and feedforward neural network. Each layer of decoder consists of three sublayers: occlusion multiplex self-attention mechanism, coding-decoder multiplex attention mechanism, and feedforward neural network.

The encoder is responsible for encoding the input traffic flow feature sequence and mapping it to the intermediate vector containing the input feature information. The decoder is responsible for decoding the intermediate vector of the encoder output into the output sequence, which involves the masked multihead self-attention.

The tag data are the input of the masked multihead self-attention module, and the dependency between tag data can be attained through the multihead self-attention mechanism [43]. The dependence is inputted into the encoding and decoding multiattentional mechanism module so that the prediction model can comprehensively learn the dependence between input feature vectors, or between label data and their dependence on each other.

The feedforward neural network is composed of two layers of neurons. The activation function of the first layer is ReLU, and the activation function of the second one is identity. If the input of ReLU activation function is greater than 0, the return value is the value provided by the input; if the input is 0 or less, the return value is 0. The node input of

the identity activation function is equal to the node output. The expressions of the two activation functions are shown in the following equations:

$$ReLU(x) = \begin{cases} x, x > 0, \\ 0, \text{others}, \end{cases} \quad (5)$$

$$Identity(x) = x. \quad (6)$$

4. Data Sources and Evaluation Criteria

4.1. Data Source. In this paper, the flow data are taken from a freeway toll station in Guangdong Province, and the statistical time is from 8 : 00 am on August 1, 2018, to 8 : 00 am on September 1, 2018. The data were counted at 15 minutes interval. A total of 2976 pieces of data were collected, as shown in Figure 2. Intuitively, the data have a certain periodicity, and the deep learning algorithm can better learn the feature changes in the sequential data [44].

In terms of data processing, these data were firstly sliced by four-step sizes to obtain a tensor of $2972 * 4$ dimensions as input data. Each step is of 15 minutes interval. The input data are divided into training sets and test sets. There are 2600 training sets and 372 test sets.

4.2. Evaluation. The evaluation criteria in this paper are mean absolute error (MAE), root-mean-square error (RMSE), and mean absolute percentage error (MAPE).

MAE error calculation is shown as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\widehat{y}_i - y_i|. \quad (7)$$

The RMSE error calculation is as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\widehat{y}_i - y_i)^2}. \quad (8)$$

The MAPE error calculation is according to the following equation:

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\widehat{y}_i - y_i}{y_i} \right|, \quad (9)$$

where n is the total number of samples, $\widehat{y}_i$ is the predicted result, and y_i is the true value.

5. Experimental

To ensure the feasibility and repeatability, equipment and parameters of the experiment were explained. Tensor-Flow2.2 deep learning framework in Anaconda environment in Windows10 system is used in this work. The CPU is Intel(R) i5-8300h, 4-core processor, 2.30 GHz, and 8 GB memory. GPU is NVIDIA GTX1050Ti, 4 GB memory, CUDA version is 10.1, CUDNN version is 7.6.
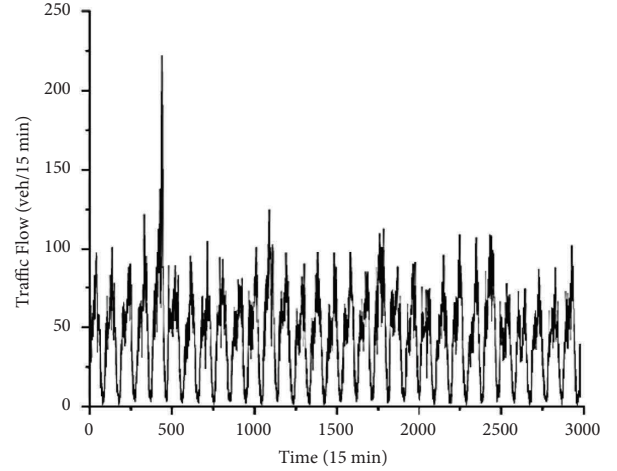

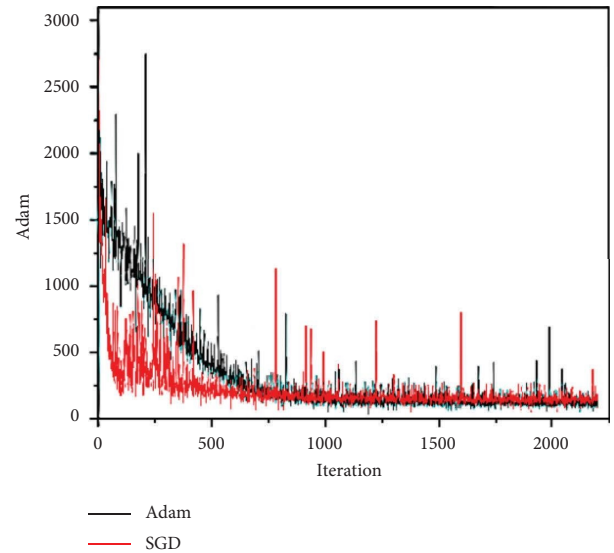Figure 2: Highway toll 15 minutes interval traffic flow.

Figure 3: The comparison of training loss curves between two optimizers.

5.1. Batch Size Setting. Batch size represents the number of samples required for a single training, and its size affects the optimization degree and speed of the model, as well as memory usage. Since the setting of a different number of training samples will lead to overfitting of data, this work sets batch sizes of 64, 128, 256, and 512 during the training model. After the preliminary test, when the batch size is 256, the training speed can be guaranteed, and a relatively low training and testing error can be obtained.

5.2. Optimizer. The optimizer is used to update and calculate network parameters that affect model training and output to approximate or reach optimal values, thereby minimizing (or maximizing) the loss function. By comparing SGD and Adam optimizers, the obtained training loss curve is shown in Figure 3. It can be seen from the figure that the convergence speed of the SGD optimizer is faster than that of the Adam optimizer, while the final error loss is similar.

The final experimental hyperparameter settings are shown in Table 1.

5.2.1. Description of Hyperparameters. The step size of the gradient descent is the learning rate, and the function is to scale the gradient. The learning decay rate is the scaling of the learning rate. Its function is to prevent overfitting and make the prediction result close to the optimal solution. The number of iterations is the number of times the network updates a parameter.

5.3. Loss Function. In this paper, the MSE loss function is adopted, which will change the gradient direction with the change of error and make the model easier to approach the optimal solution.

6. Results and Discussion

To verify the effectiveness of the proposed model on predicting traffic flow, this paper selects three baseline models for comparison. The performance of the proposed model (STGGA) is compared with SVR, LSTM-attention, and TCN models.

SVR: SVR is a popular machine learning approach for the short-term traffic flow prediction [27, 45]. Consequently, a basic SVR model with a linear kernel is selected as one of our baselines.

LSTM-attention: LSTM-attention is a LSTM model with an attention module [36, 46]. The attention mechanism module is added to the LSTM model to deal with the problem of gradient disappearance in the traffic flow data of long-term series, while LSTM-attention contains two LSTM layers and connects to the attention mechanism. So, the model is selected as one of our baselines.

TCN: TCN is a residual structure neural network model based on full convolution to solve the problems of gradient stability and inability to process the information in parallel in RNN [37, 47], while TCN contains two one-dimensional convolution layers and one one-dimensional pooling layer. Thus, the model is selected as one of our baselines.

The evaluation indexes of the prediction results for each model are listed in Table 2. As listed in Table 2, MAE, RMSE, and MAPE errors of the proposed model in our work are all smaller than those of the comparison models. Since traffic flow observations vary from a few hundred vehicles per hour in off peak to several thousand vehicles during peak periods, a MAPE in the range of 10–20% is generally acceptable in most studies on the flow prediction [10, 48, 49].

First, the performance of the proposed model is compared with the SVR model. In the literature [50], SVR predicted MAPE results, that is, 15.71%. The prediction result of this paper is 21.7%. Different from the dataset used in our work, the dataset in the literature is the traffic data of an intersection in Jinan. The data used in this paper are from freeway Toll station and are more volatile, so the MAPE value in this paper is different from the predicted value in the literature. Compared with the SVR model, the RMSE and MAE error of the proposed model is reduced by 0.9 and 0.89,

respectively, and the MAPE error of the proposed model is reduced by 8.18%. Because SVR uses a kernel function to map a large number of uncertain traffic flow data to high-dimensional space, it cannot make full use of spatial and periodic characteristics for prediction, thus resulting in limited predictive performance.

Next, performance of the proposed model is compared with the TCN method. In the literature [51], TCN-predicted MAPE results were various as 17.72%, 21.87%, 15.4%, and 19.45%. The prediction results of this paper are in a normal range. Compared with the TCN model, RMSE and MAE errors of the proposed model are reduced by 0.84–0.08, respectively, and the MAPE error of the proposed model is reduced by 4.95% due to the limited receptive field size of convolutional neural networks for long-term sequence problems.

Furthermore, the proposed model is compared with the LSTM-attention model at last to verify the effect of self-attention mechanism model. In previous literature [52], MAPE predicted by LSTM-attention is 15.79%, while it is 16.08% in our work. Compared with the LSTM-attention model, RMSE and MAE errors of the proposed model is reduced by 0.42–0.36, respectively, and the MAPE error of the proposed model is reduced by 3.09%. The results demonstrate that the proposed model is more accurate than the LSTM-attention.

The box diagram of model error is shown in Figure 4. The median error of the proposed model is 3.17, and the error of the compared baseline model is 6.78 for SVR, 5.59 for LSTM-attention, and 6.59 for TCN. From the upper and lower ranges of prediction error, the error range of the proposed model is smaller and the prediction result is more stable. In terms of outliers, 1.5 times the difference between the upper and lower quartiles (IQR) is used as the judgment standard. The results show that among the 372 test set data, the prediction outliers of the four models are less than 5%, and there is no large deviation in the prediction results.

To illustrate the fitting results clearly, the fitted curve volume from these models is shown in Figure 5. According to the fitted curve in Figure 5(a), the prediction results of the model proposed in this paper are much closer to the real value and are extremely excellent. The prediction effect of SVR model is the worst of all. It can be seen from Figure 5(b) that there is a certain lag in the prediction results compared with the real value. Compared with SVR model, the LSTM-attention model and TCN model have relatively higher accuracy, but still with a certain hysteresis. As shown in Figure 5(c), this lag may be affected by data autocorrelation. As illuminated in Figure 5(d), the prediction of the proposed model is more effective in dealing with this lag. In contrast, the model proposed in our work can be screened and featured for critical information over a long-term period, thus better predictive effects can be achieved.

The main values of this work are shown as follows: The prediction effect of the proposed model is better for the peak and low periods of traffic volume compared to baseline models. It indicates that the prediction performance of the proposed model for extreme conditions is

TABLE 1: Hyperparameters of the experiment.

Initial learning rate	Learning decay rate	Training epochs	Optimizer	Batch size
$3 * 10^{-4}$	0.5	200	SGD	256

TABLE 2: The experiment result comparison.

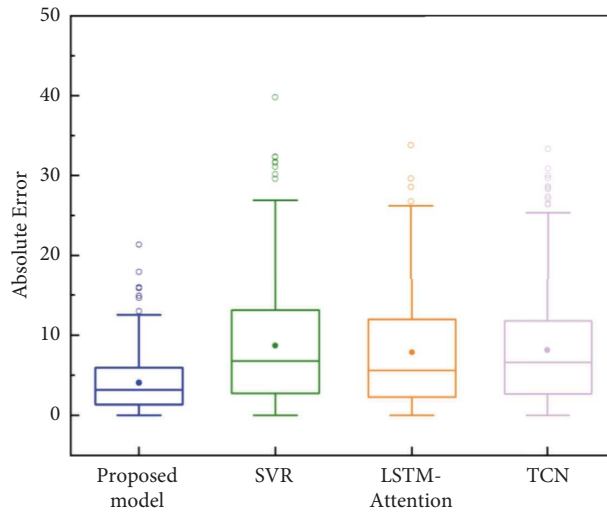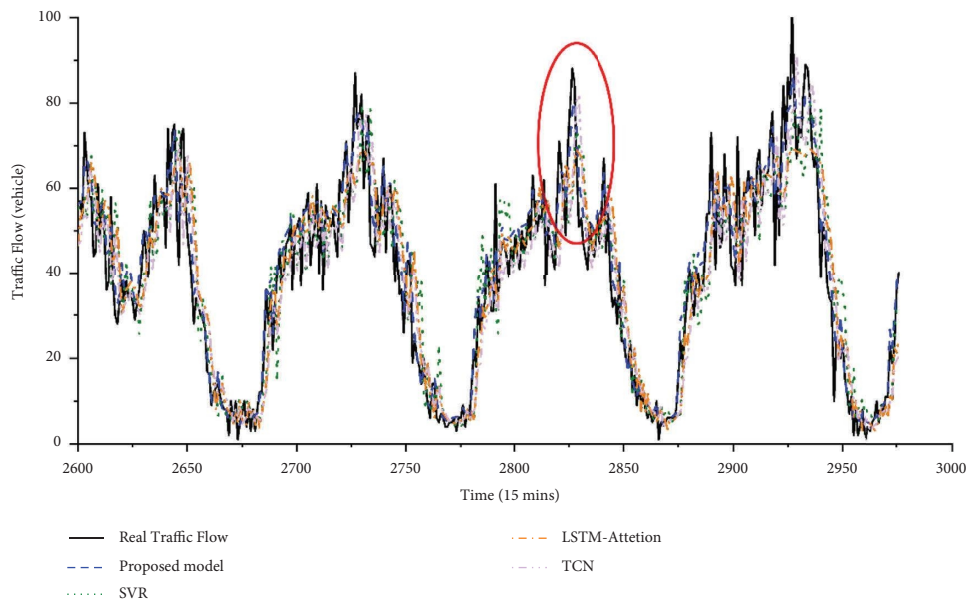| Criteria	Proposed model	SVR	LSTM-attention	TCN
RMSE	3.01	3.90	3.37	3.89
MAE	4.38	5.28	4.80	5.22
MAPE	12.99%	21.17%	16.08%	17.94%

FIGURE 4: The box chart comparison.
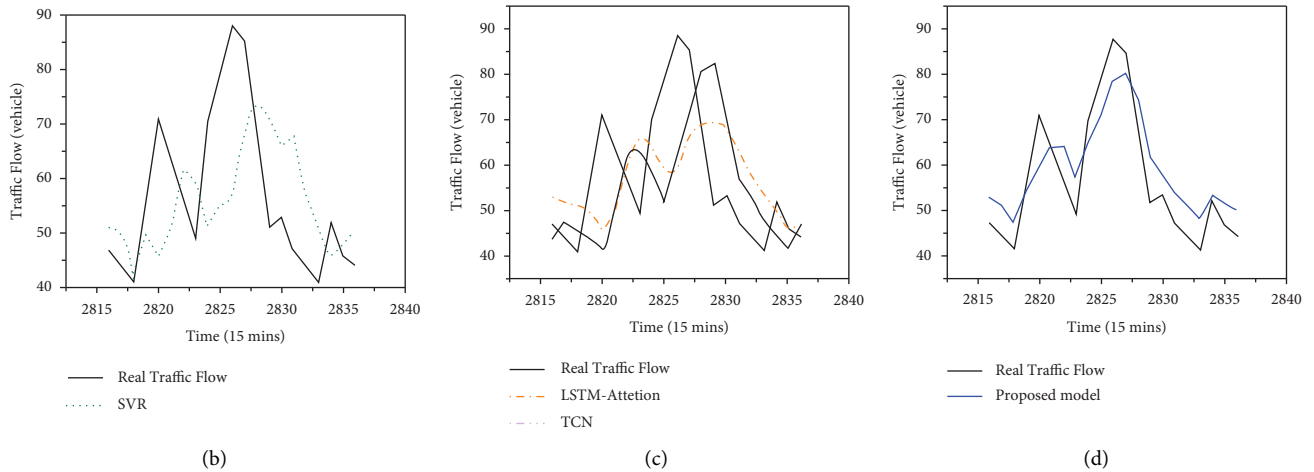

(a)

FIGURE 5: Continued.

FIGURE 5: The traffic flow prediction comparison. (a) The comparison diagram of the original traffic flow with the proposed model, SVR model, LSTM-attention model, and TCN model; (b) the comparison diagram of the SVR model predicted the traffic flow with real traffic volume; (c) the comparison diagram of the LSTM-attention model with the TCN model and real traffic volume; (d) true time intersection comparison of flux and proposed model predicted flux.

quite good. Moreover, the model is well adapted to the characteristics of traffic flow fluctuations caused by spatial relationships, weather, vacation, and other factors [53, 54]. In the future, these factors can be further refined to improve the prediction accuracy of the model. It is very effective to add attention mechanisms to improve the hybrid model performance [36, 55, 56]. But its interpretability is still questioned. The interpretability of the attention mechanism played an important role in the proposed model to solve the problems, which is mentioned in the literature [57].

7. Conclusions

It is very important to reduce traffic carbon emissions for the living environment. Vehicle exhaust is one of the main sources of carbon emissions. The proposed model for the short-term traffic flow prediction is very effective to improve the prediction accuracy.

In this work, aiming at the gradient disappearance or gradient explosion and information dissipation in long-term series in short-term traffic flow prediction, a novel self-attention-based hybrid model is proposed to enhance the extraction ability of key information. The experimental results show that the proposed model can achieve a more accurate prediction effect in the case of multiple traffic flow fluctuations. Compared with several conventional time series prediction models, the proposed model in this paper has less error and is much closer to the real value.

Certainly, the modified model in this paper has some limitations and further research should pay close attention to (1) the influence factors such as weather and space characteristics should be considered to increase the generalization ability and further improve the prediction accuracy of the model. (2) The interpretability of deep learning models is insufficient and needs further strengthen.

(3) The performance of the model for complex data with a large sample size remains to be verified.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Additional Points

Highlights. (i) Novel self-attention-based hybrid model for short-term traffic flow prediction is proposed. (ii) The encoder-decoder framework is developed to extract dominant spatiotemporal features and short-term variations. (iii) Three well-known benchmarks are considered to evaluate the models. (iv) The proposed model showed better results than SVR, LSTM, LSTM-attention, and TCN according to the real case.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Zhihong Li and Xiaoyu Wang conceptualized and designed the study. Xiaoyu Wang, Yang Dong, and Kairan Yang collected data. Zhihong Li and Xiaoyu Wang analyzed and interpreted data. Zhihong Li prepared draft manuscript. All authors reviewed the results and approved the final version of the manuscript.

Acknowledgments

References

[1] B. Ghosh, B. Basu, and M. O'Mahony, "Multivariate short-term traffic flow forecasting using time-series analysis," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, no. 2, pp. 246–254, 2009.

[2] M. J. Al Sayah, C. Abdallah, M. Khouri, R. Nedjai, and T. Darwich, "A framework for climate change assessment in Mediterranean data-sparse watersheds using remote sensing and ARIMA modeling," *Theoretical and Applied Climatology*, vol. 143, no. 1-2, pp. 639–658, 2021.

[3] K. Yunus, T. R. Thiringer, and P. Chen, "ARIMA-basedfrequency-decomposed modeling of wind speed time series," *IEEE Transactions on Power Systems*, vol. 31, no. 4, pp. 2546–2556, 2016.

[4] H. Zheng, F. Lin, X. Feng, and Y. Chen, "A hybrid deep learning model with attention-basedconv-LSTM networks for short-term traffic flow prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 11, pp. 6910–6920, 2021.

[5] M. W. Li, W. C. Hong, and H. G. Kang, "Urban traffic flow forecasting using Gauss–SVR with cat mapping, cloud model and PSO hybrid algorithm," *Neurocomputing*, vol. 99, pp. 230–240, 2013.

[6] W. C. Hong, Y. Dong, F. Zheng, and S. Y. Wei, "Hybrid evolutionary algorithms in a SVR traffic flow forecasting model," *Applied Mathematics and Computation*, vol. 217, no. 15, pp. 6733–6747, 2011.

[7] M. Lopez-Martin, B. Carro, A. Sanchez-Esguevillas, and J. Lloret, "Network traffic classifier with convolutional and recurrent neural networks for internet of things," *IEEE Access*, vol. 5, Article ID 18042, 18050 pages, 2017.

[8] X. Ma, H. Zhong, Y. Li, J. Ma, Z. Cui, and Y. Wang, "Forecasting transportation network speed using deep capsule networks with nested LSTM models," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 8, pp. 4813–4824, 2021.

[9] A. Kumar Dubey, A. Kumar, V. García-Díaz, A. Kumar Sharma, and K. Kanhaiya, "Study and analysis of SARIMA and LSTM in forecasting time series data," *Sustainable Energy Technologies and Assessments*, vol. 47, Article ID 101474, 2021.

[10] S. Siami-Namini, N. Tavakoli, and A. Siami Namin, "A Comparative Analysis of Forecasting Financial Time Series Using ARIMA, LSTM, and BiLSTM," p. 9512, 2019, https://arxiv.org/abs/1911.09512.

[11] Z. Li, H. Xu, X. Gao, Z. Wang, and W. Xu, "Fusion attention mechanism bidirectional LSTM for short-term traffic flow prediction," *Journal of Intelligent Transportation Systems*, pp. 1–14, 2022.

[12] D. Ienco, Y. J. E. Gbodjo, R. Gaetano, and R. Interdonato, "Weakly supervised learning for land cover mapping of satellite image time series via attention-based CNN," *IEEE Access*, vol. 8, Article ID 179547, 179560 pages, 2020.

[13] F. Kong, J. Li, B. Jiang, H. Wang, and H. Song, "Integrated generative model for industrial anomaly detection via Bi-directional LSTM and attention mechanism," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 1, pp. 541–550, 2023.

[14] Y. M. Wang, Z. Zhang, L. W. Feng, Y. C. Ma, and Q. Y. Du, "A new attention-based CNN approach for crop mapping using time series Sentinel-2 images," *Computers and Electronics in Agriculture*, vol. 184, Article ID 106090, 2021.

[15] X. F. Yuan, L. Li, Y. A. W. Shardt, Y. L. Wang, and C. H. Yang, "Deep learning with spatiotemporal attention-based LSTM for industrial soft sensor model development," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 5, pp. 4404–4414, 2021.

[16] H. Yan, X. Ma, and Z. Pu, "Learning dynamic and hierarchical traffic spatiotemporal features with transformer," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 22386–22399, 2022.

[17] W. Tao, C. Li, R. Song et al., "EEG-based emotion recognition via channel-wise attention and self attention," *IEEE Transactions on Affective Computing*, pp. 1–14, 2020.

[18] S. Shahriari, M. Ghasri, S. A. Sisson, and T. Rashidi, "Ensemble of ARIMA: combining parametric and bootstrapping technique for traffic flow prediction," *Transportmetrica: Transportation Science*, vol. 16, no. 3, pp. 1552–1573, 2020.

[19] M. Van Der Voort, M. Dougherty, and S. Watson, "Combining Kohonen maps with ARIMA time series models to forecast traffic flow," *Transportation Research Part C: Emerging Technologies*, vol. 4, no. 5, pp. 307–318, 1996.

[20] L. Shen, J. Lu, D. Geng, and L. Deng, "Peak traffic flow predictions: exploiting toll data from large expressway networks," *Sustainability*, vol. 13, no. 1, p. 260, 2020.

[21] H. Chao, S. Su, and W. Cheng-hong, "A real-timeshort-term traffic flow adaptive forecasting method based on ARIMA model," *Journal of System Simulation*, vol. 16, no. 7, pp. 1530–1535, 2004.

[22] B. M. Williams and L. A. Hoel, "Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: theoretical basis and empirical results," *Journal of Transportation Engineering*, vol. 129, no. 6, pp. 664–672, 2003.

[23] Y. Kamarianakis and P. Prastacos, "Forecasting traffic flow conditions in an urban network: comparison of multivariate and univariate approaches," *Transportation Research Record*, vol. 1857, no. 1, pp. 74–84, 2003.

[24] W. Xiaoquan, S. Chunfu, Y. Chaoying, J. Xun, and G. Ling, "Short term traffic flow forecasting method based on ARIMA-GARCH-M model," *Journal of Beijing Jiaotong University*, vol. 42, no. 4, pp. 79–84, 2018.

[25] A. Cheng, X. Jiang, Y. Li, C. Zhang, and H. Zhu, "Multiple sources and multiple measures based traffic flow prediction using the chaos theory and support vector regression method," *Physica A: Statistical Mechanics and Its Applications*, vol. 466, pp. 422–434, 2017.

[26] C. Luo, C. Huang, J. Cao et al., "Short-term traffic flow prediction based on least square support vector machine with hybrid optimization algorithm," *Neural Processing Letters*, vol. 50, no. 3, pp. 2305–2322, 2019.

[27] X. Feng, X. Ling, H. Zheng, Z. Chen, and Y. Xu, "Adaptive multi-kernel SVM with spatial-temporal correlation for short-term traffic flow prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 6, pp. 2001–2013, 2019.

[28] D. Wei and H. Liu, "An adaptive-margin support vector regression for short-term traffic flow forecast," *Journal of Intelligent Transportation Systems*, vol. 17, no. 4, pp. 317–327, 2013.

[29] Z. Li, Y. Dong, Y. Wen, H. Xu, and J. Wu, "A deep pedestrian tracking SSD-based model in the sudden emergency or violent environment," *Journal of Advanced Transportation*, vol. 2021, Article ID 2085876, 13 pages, 2021.

[30] Y. Chen, Y. Kang, Y. Chen, and Z. Wang, "Probabilistic forecasting with temporal convolutional neural network," *Neurocomputing*, vol. 399, pp. 491–501, 2020.

[31] B. Zhao, H. Lu, S. Chen, J. Liu, and D. Wu, "Convolutional neural networks for time series classification," *Journal of*

Systems Engineering and Electronics, vol. 28, no. 1, pp. 162–169, 2017.

[32] Z. Cui, R. Ke, Z. Pu, and Y. Wang, "Deep Bidirectional and Unidirectional LSTM Recurrent Neural Network for Network-wide Traffic Speed Prediction," 2018, https://arxiv.org/abs/1801.02143.

[33] L. Cai, M. Lei, S. Zhang, Y. Yu, T. Zhou, and J. Qin, "A noise-immune LSTM network for short-term traffic flow forecasting," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 30, no. 2, Article ID 23135, 2020.

[34] C. Ma, G. Dai, and J. Zhou, "Short-term traffic flow prediction for urban road sections based on time series analysis and LSTM_BILSTM method," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 5615–5624, 2022.

[35] W. Wei, H. Wu, and H. Ma, "An autoencoder and LSTM-based traffic flow prediction method," *Sensors*, vol. 19, no. 13, p. 2946, 2019.

[36] B. L. Yang, S. L. Sun, J. Y. Li, X. X. Lin, and Y. Tian, "Traffic flow prediction using LSTM with feature enhancement," *Neurocomputing*, vol. 332, pp. 320–327, 2019.

[37] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: a unified approach to action segmentation," in *European Conference on Computer Vision*-Springer, Berlin, Germany, 2016.

[38] Z. Cui, K. Henrickson, R. Ke, and Y. Wang, "Traffic graph convolutional recurrent neural network: a deep learning framework for network-scale traffic learning and forecasting," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 11, pp. 4883–4894, 2020.

[39] M. Lv, Z. Hong, L. Chen, T. Chen, T. Zhu, and S. Ji, "Temporal multi-graph convolutional network for traffic flow prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 6, pp. 3337–3348, 2020.

[40] Y. Qiao, Y. Wang, C. Ma, and J. Yang, "Short-term traffic flow prediction based on 1DCNN-LSTM neural network structure," *Modern Physics Letters B*, vol. 35, no. 2, Article ID 2150042, 2021.

[41] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: data-driven traffic forecasting," 2018, https://arxiv.org/abs/1707.01926.

[42] K. Cho, B. Van Merriänboer, C. Gulcehre et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, https://arxiv.org/abs/1406.1078.

[43] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[44] Z. Shen, Y. Zhang, J. Lu, J. Xu, and G. Xiao, "A novel time series forecasting model with deep learning," *Neurocomputing*, vol. 396, pp. 302–313, 2020.

[45] M. Castro-Neto, Y. S. Jeong, M. K. Jeong, and L. D. Han, "Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions," *Expert Systems with Applications*, vol. 36, no. 3, pp. 6164–6173, 2009.

[46] D. Chen, C. Xiong, and M. Zhong, "Improved LSTM based on attention mechanism for short-term traffic flow prediction," in *Proceedings of the 2020 10th International Conference on Information Science and Technology (ICIST)*, Bath, London, September 2020.

[47] S. Liu, P. Cui, W. Zhu, and S. Yang, "Learning socially embedded visual representation from scratch," in *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 109–118, Brisbane, Australia, October 2015.

[48] S. V. Kumar and L. Vanajakshi, "Short-term traffic flow prediction using seasonal ARIMA model with limited input data," *European Transport Research Review*, vol. 7, no. 3, p. 21, 2015.

[49] D. Yanjie, L. Yisheng, and W. Fei-Yue, "Travel time prediction with LSTM neural network," in *Proceedings of the 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, Rio de Janeiro, Brazil, November 2016.

[50] W. Li, J. Wang, R. Fan et al., "Short-term traffic state prediction from latent structures: accuracy vs. efficiency," *Transportation Research Part C: Emerging Technologies*, vol. 111, no. 1, pp. 72–90, 2020.

[51] X. Xian-feng, Y. Fan, L. Zhuang-zhuang, L. Long-jie, and X. Zhen, "Research on the combination model of short-term traffic flow prediction based on two-level screening mechanism," *Control and Decision*, vol. 38, no. 1, pp. 1–9, 2021.

[52] W. Qingring, T. Keke, Z. Changfeng, and W. Yimeng, "Short term traffic flow prediction based on multi factors," *Computers Engineering and Applications*, vol. 58, no. 21, pp. 1–12, 2022.

[53] X. Xu, Z. Fang, L. Qi, X. Zhang, Q. He, and X. Zhou, "Tripres: traffic flow prediction driven resource reservation for multimedia iov with edge computing," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 17, no. 2, pp. 1–21, 2021.

[54] Y. Jia, J. Wu, and M. Xu, "Traffic flow prediction with rainfall impact using a deep learning method," *Journal of Advanced Transportation*, vol. 2017, Article ID 6575947, 10 pages, 2017.

[55] H. Lu, Z. Ge, Y. Song, D. Jiang, T. Zhou, and J. Qin, "A temporal-aware LSTM enhanced by loss-switch mechanism for traffic flow forecasting," *Neurocomputing*, vol. 427, pp. 169–178, 2021.

[56] K. Wang, C. X. Ma, Y. H. Qiao, X. J. Lu, W. N. Hao, and S. Dong, "A hybrid deep learning model with 1DCNN-LSTM-Attention networks for short-term traffic flow prediction," *Physica A: Statistical Mechanics and its Applications*, vol. 583, Article ID 126293, 2021.

[57] S. Jain and B. C. Wallace, "Attention is not explanation," 2019, https://arxiv.org/abs/1902.10186.