Vehicle Trajectory Prediction Model Based on Fusion Neural Network

Xuemei Mou(✉), Xiang Yu, Binbin Wang, Ziyi Wang, and Fugui Deng

School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China
mouxuemei175101@163.com

Abstract. To address the problem of the lack of interpretability of vehicle trajectory prediction models based on deep learning, this paper proposes a Fusion Neural network with the Spatio-Temporal Attention (STA-FNet) model. The model outputs a predictive distribution of future vehicle trajectories based on different vehicle trajectories and traffic environment factors, with an in-depth analysis of the Spatio-temporal attention weights learned from various urban road traffic scenarios. In this paper, the proposed model is evaluated using the publicly available NGSIM dataset, and the experimental results show that the model not only explains the influence of historical trajectories and road traffic environment on the target vehicle trajectories but also obtains better prediction results in complex traffic environments.

Keywords: Intelligent traffic · Trajectory prediction · Deep learning · Spatial-temporal relationships

1 Introduction

Urban traffic congestion is becoming increasingly serious, and accurate prediction of vehicle trajectories is crucial for urban intelligent transportation. Currently, most trajectory predictions rely mainly on historical traffic data to predict and find rules from temporal features [1–3]. In recent years, with the development of neural network models, more and more scholars use neural networks as trajectory prediction models. Yang et al. [4] used Long Short Term Memory (LSTM) to predict the trajectory of the preceding vehicle with good results by joint time series modeling of vehicles with different driving styles around the target vehicle. Kaushik et al. [5] used Recurrent Neural Network (RNN) model to analyze the real-time series data acquired by in-vehicle sensors and the model showed good performance in predicting the future trajectory of an obstacle vehicle.

In complex data prediction, it is often difficult for a single model to capture the complexity and variability of trajectory data at the same time, failing to maintain good prediction performance, while combined models have good results. Ip et al. [6] utilize a combined model of LSTM and RNN to predict vehicle trajectories in the case the current

F. Gao et al. (Eds.): ChinaCom 2022, LNICST 500, pp. 309–320, 2023.
https://doi.org/10.1007/978-3-031-34790-0_24

and previous positions of the vehicle are known. Rossi et al. [7] used a combined model of LSTM and Generative Adversarial Network (GAN) to predict vehicle trajectories and performed well in scenes with high multimodal effects. Based on the data collected by the sensor, Wang et al. [8] used the combined model of Convolutional Neural Network (CNN) and RNN to conduct an in-depth analysis of the vehicle motion data collected by the sensor, to achieve the purpose of protecting the scene texture information of the environment and the interaction between the surrounding vehicles.

The attention mechanism proposed by Bahdanau et al. [9] can be naturally integrated with RNN to improve model interpretability. Cai et al. [10] used Graph Convolution Network (GCN) to pay attention to the interaction between the vehicle and the non-Euclidean-related structures existing in the environment and used the attention mechanism to enhance the extraction of image features by GCN. Yu et al. [11] managed the importance of the driving flow of the target and neighboring vehicles and the dynamics of the target vehicle in each driving situation by using an attention mechanism and utilizing LSTM to predict future trajectories.

To sum up, most of the related research focuses on the prediction of the vehicle's motion state and historical trajectory, while the interaction between the target vehicle and the road traffic environment, the spatial position of the surrounding vehicles, and the interaction with the target vehicle still need further research. Taking this as a motivation, this paper proposes a Fusion Neural Network with Spatio-Temporal Attention (STA-FNet) model through the analysis of the real urban vehicle running state data set. To predict the trajectory, it is proposed to extract the Spatio-temporal relationship between the target vehicle, surrounding vehicles, and road environment information through a model combined with a Convolutional Social pooling (CS) and a Bidirectional Recurrent Gating Unit (BiGRU). The main contributions of this paper are as follows:

1. Combination of BiGRU and CS, a novel and robust vehicle trajectory prediction structure are proposed. The model can not only explain the Spatio-temporal features between vehicle trajectory data but also quantify these factors for trajectory prediction.
2. The convolutional social pooling layer captures the target vehicle trajectory data and the spatial relationship between the target vehicle and surrounding vehicle trajectory data, and introduces a spatial attention mechanism to increase the extraction of key influencing factors.
3. BiGRU instead of traditional RNN can not only fully consider the Spatio-temporal relationship between data, but also make up for the defect that CS cannot effectively extract long-term sequence features. And through the attention module for mapping weighting and learning parameter matrix to give different weights to the hidden state of BiGRU to further improve the prediction accuracy of the model.

The organization of this paper is as follows: Sect. 2 related definitions and questions; Sect. 3 discusses the work related to the research in this paper; Sect. 4 describes the proposed method and models it; Sect. 5 experimental results; Sect. 6 summarizes the related work done in this paper.

2 Problem Analysis

First, in a static traffic scene at a certain moment, the eigenvalues (parking space, vehicle speed, acceleration, etc.) of the historical trajectory of the target vehicle at the current moment in various states will affect the future trajectory of the vehicle. Second, the historical spatial positions of the surrounding vehicles and the interaction with the target vehicle also affect the future trajectory of the vehicle. Therefore, this paper establishes a local reference frame for the predicted scene, making the model independent of the curvature of the road. The origin of the prediction at time t is on the target vehicle, as shown in Fig. 1, the y-axis points to the direction of movement of the road, and the x-axis points to the direction perpendicular to the road. T is the observation period.

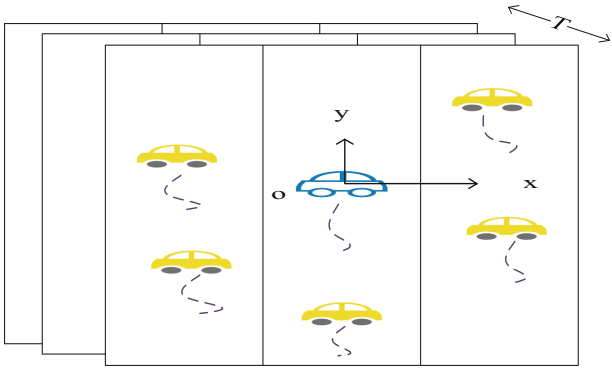

Fig. 1. Reference coordinate system.

Combined with the above content, this paper will initialize the state features contained in the historical trajectories of all vehicles in the traffic scene at the current moment. The historical trajectory state feature of the k-th vehicle is expressed as:

$$\tilde{x}_k = \{x, y, v, a, d, c\} \tag{1}$$

$$\tilde{x}_k^{(t-t_i)} = x_k^{(t-t_i)} - x_k^{(t)} \tag{2}$$

Among them, $t_i \in T$, $k \in n$, and n is the number of all vehicles in the current traffic scene, and T is the time length of the historical trajectory. (x, y) is the position coordinates of the above-relativized vehicles, respectively. v, a, and d are the vehicle speed, acceleration, and the relative distance between the surrounding vehicles and the target vehicle respectively. c is the road congestion coefficient index at the current moment. X_k is the historical trajectory feature sequence of the vehicle, which is expressed as:

$$X_k = \left\{ \overset{\sim(t-T)}{x_k}, \cdots, \overset{\sim(t-1)}{x_k}, \overset{\sim(t)}{x_k} \right\} \tag{3}$$

3 Related Theory

3.1 Convolutional Social Pooling

CNN adopts the method of weight sharing, which has advantages in mining local relevant information in space. While convolving the trajectory vectors of the surrounding vehicles, to improve the accuracy of the local position of the vehicle, the influence of the existence of the target vehicle on the decision-making of the surrounding vehicles is also considered. In this paper, CS is used to learn the interdependence in the process of vehicle motion more robustly, to analyze the upstream and downstream spatial relationship between the vehicle trajectory data, extract multiple key features, and select the ReLU for activation.

$$C_1 = f(X \otimes W_1 + b_1) = ReLU(X \otimes W_1 + b_1) \tag{4}$$

$$C_2 = f(P_1 \otimes W_2 + b_2) = ReLU(P_1 \otimes W_2 + b_2) \tag{5}$$

$$P_1 = \max\left({C_2}^{\mathrm{T}}\right) + b_3 \tag{6}$$

$$h_c = {P_1}^{\mathrm{T}} \tag{7}$$

where C_1 and C_2 is the output of convolutional layer 1 and convolutional layer 2 respectively, P_1 is the output of the max pooling layer, W_1, W_2 and W_3 are the weight matrices, b_1, b_2, b_3 and b_4 are the biases, $\otimes$ is the convolution operation. And the local positions between the target vehicle and neighboring vehicles, the interaction between the target vehicle and the environment and the spatial relationships between the trajectory data are captured by the CS network in $(t - T), \cdots, (t - 1), t$ period. The output feature vector H_c can be expressed as:

$$H_c = \left[h_{c_1}, h_{c_2}, \cdots, h_{c_i}\right]^T \tag{8}$$

3.2 Prediction Model

After multi-feature extraction is performed on the observation data of the vehicle running state, the observation data sequence is represented as a vector matrix. When the traditional RNN network processes time series information, the problem of gradient explosion or gradient disappearance occurs, as the length of the time series increases. Considering the advantage that Bidirectional GRU can process both forward and backward information of long-term series at the same time, this paper chooses BiGRU as the prediction model. It is a double-layer structure composed of two GRU in different directions. Trajectory sequences can provide not only forward information but also backward derivation references if known below. At time t, the specific calculation formula of BiGRU is as follows:

$$\overleftarrow{h}_t = GRU\left(W_t, \overleftarrow{h}_{t-1}\right), t \in [1, m] \tag{9}$$

$$\overleftarrow{h}_t = GRU\left(W_t, \overleftarrow{h}_{t-1}\right), t \in [1, m] \tag{10}$$

$$h_t = [\overrightarrow{h}_t, \overleftarrow{h}_t] \tag{11}$$

The final output of BiGRU is denoted as:

$$H_B = [h_1, h_2, \cdots, h_t] \tag{12}$$

3.3 Attention Mechanism

To improve the efficiency of trajectory prediction, this paper uses the Soft Attention model proposed by Bahdanau et al. [9], in which the degree of attention of each information area is represented by a weighted score in the range of [0,1]. The calculation formula of the weight coefficient of the Attention mechanism layer can be expressed as:

$$e_t = \mu tan\mathrm{h}(w\mathrm{h}_t + b) \tag{13}$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_{j-1}^{t} e_j} \tag{14}$$

$$s_t = \sum_{t=1}^{i} \alpha_t h_t \tag{15}$$

Among them, e_t represents the attention vector at time t, u and w are the weight coefficients, b is the bias coefficient; α_t is the attention weight; the context vector s_t of each data will be calculated, that is, the output of this layer.

4 Model Building

4.1 Model Frame

According to the definition and description in Sect. 2 of this paper, the input of the STA-FNet model is the historical trajectories of all vehicles in the $(t-T), \cdots, (t-1), t$ period, the motion state information, and road congestion status information. The input matrix is represented as:

$$X_T = \{X_1, \cdots X_k, \cdots X_n\} \tag{16}$$

To formally express the problem to be solved in this paper, let $P_{X_T}(X)$ denote the probability that the trajectory X appears in the trajectory X_T. Based on the definition above, the method proposed in this paper can obtain the trajectory set that maximizes $P_{X_T}(X)$. A more explicit definition is as follows:

$$maxP_{X_T}(X_1, \cdots X_k, \cdots X_n) \rightarrow \mathrm{Y} = \{X_1^{(t+t_m)}, \cdots, X_k^{(t+t_m)}, \cdots X_n^{(t+t_m)}\} \tag{17}$$

4.2 Model Implementation

To comprehensively consider the influence of surrounding vehicles and road traffic environment on driving trajectories, a fusion neural network model with Spatio-temporal attention is built in this paper. Firstly, the road network is modeled, and the road congestion condition at the current moment is one-hot encoded so that the encoded vector can better reflect the dynamic change of road congestion conditions over time. And the local positions between the target vehicle and neighboring vehicles, the interaction between the target vehicle and the environment, and the spatial relationships between the trajectory data are captured by the convolutional social pooling (CS) network in $(t-T), \cdots, (t-1), t$ period. Meanwhile, spatial attention acts to increase the extraction of key features. Secondly, after training through the CS network the above key features are represented as vector matrices, and the Spatio-temporal relationships between the trajectory data are continued to be extracted and fused by the two-layer BiGRU network to make reasonable predictions of future trajectories and filter out the vehicle trajectory sequences with the highest probability. Meanwhile, temporal attention acts to adjust the weight coefficients to extract the key features affecting the vehicle trajectory. The system framework is shown in Fig. 2:

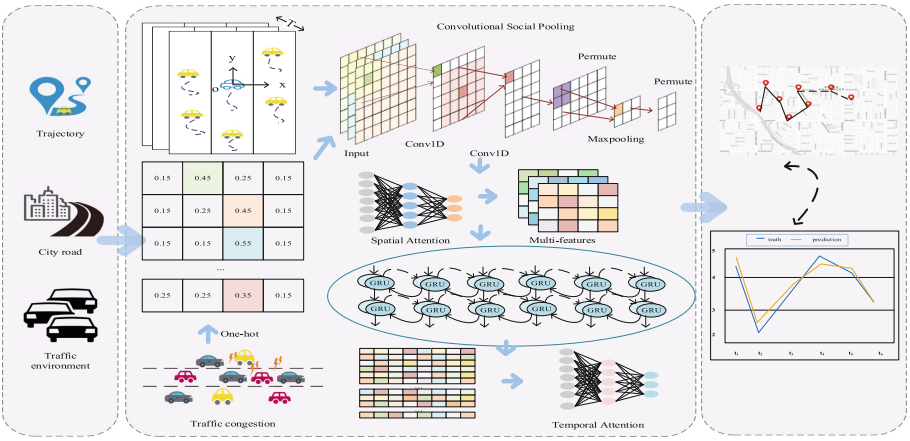

Fig. 2. System Model Framework.

4.3 Algorithm Complexity Analysis

In the stage of feature extraction using CS, the time complexity is $O(n^3)$; in the stage of prediction using vector matrix, the time complexity of the BiGRU model is $O(n^2)$, and the Spatio-temporal attention acts on the above models. The time complexity of the two stages of data processing is $O(n)$. Through the above analysis, the time complexity of the entire algorithm is $O(n^3) + O(n^2) + O(n) \sim O(n^3)$.

5 Experimental Results and Analysis

5.1 Dataset

This paper adopts the Lankershim Boulevard Urban Roads dataset from the NGSIM dataset, which exhibits real urban road traffic trajectories captured over a time of 30 min. The running status information of the vehicle in 5 different road sections is recorded, including information such as vehicle position, GPS coordinates, speed, acceleration, and vehicle type, as shown in Fig. 3. This paper selects the trajectory data of 800 consecutive frames as samples, and divides them according to the proportion of 62.5%, of which the first 500 sets of data are used as the training set, and the last 300 sets of data are used as the test set. The continuous variables are normalized to normalize the mean and variance of the data. The algorithm is optimized using the ADAM function with a learning rate of 0.01, and the number of samples per batch is 16.

Fig. 3. Distribution map of 5 road sections.

Due to the complex and changeable traffic operating environment in the urban road environment, the classification of traffic operating conditions is often inaccurate and there is a certain degree of ambiguity. Therefore, this paper uses the fuzzy comprehensive evaluation method of traffic flow to calculate the congestion coefficient of this road section. Referring to the "American Traffic Congestion Evaluation Index System", the Congestion Coefficient is defined as the ratio (V/C) of the actual volume of road traffic to the road capacity. When the Congestion Coefficient is in [0, 0.77], [0.78, 0.85], [0.86, 0.99], [1.0, 1.2], it is defined as "unblocked", "slightly congested", "moderately congested" and "severe congestion", associated with variables 1, 2, 3, and 4. One-hot coding is performed on discrete variables in this paper to unify the types of discrete variables and continuous variables. Therefore, after adopting the fuzzy evaluation method, the congestion coefficients of the five road sections can reflect the dynamic changes in road congestion to a certain extent, as shown in Fig. 4. By calculating the Congestion Coefficient, the current running state of the road section can be well understood, which is helpful for the subsequent prediction of the vehicle trajectory.

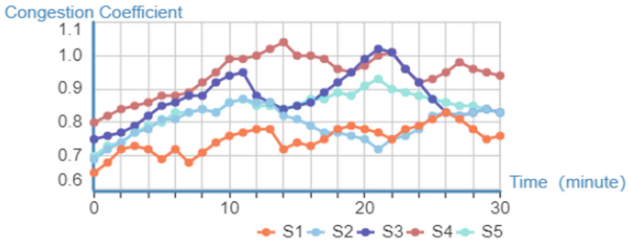

Fig. 4. Congestion Coefficient of 5 road sections.

5.2 Baseline Models

To explore the predictive performance of the proposed STA-FNet model, several baseline models are also trained and used to predict vehicle trajectories at the same time.

4. CS-LSTM [12]: Introduce a convolutional social pooling layer and LSTM to predict vehicle trajectories.
5. DCS-LSTM [13]: A dilated convolutional social pooling layer and LSTM are introduced to predict vehicle trajectories.
6. BiLSTM [14]: Predicting vehicle trajectories using Bidirectional LSTM.
7. STA-FNet: The fusion neural network prediction model based on the spatiotemporal attention mechanism proposed in this paper uses the convolutional social pooling layer to extract the spatiotemporal relationship between the historical social vectors of the target vehicle and completes the prediction through BiGRU.

5.3 Evaluation Indicators

This paper uses Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percent Error (MAPE) for evaluation. The loss function quantifies how close the neural network model is to the ideal situation it was trained on. To facilitate the calculation, using the root mean square error as the loss function of the model.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2} \tag{18}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\widehat{y_i} - y_i\right| \tag{19}$$

$$\mathrm{MAPE} = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{\widehat{y_i} - y_i}{y_i}\right| \tag{20}$$

Among them, $\widehat{y_i}$ represents the predicted value, y_i represents the true value, and N is the number of samples.

5.4 Performance Analysis

Determination of Basic Parameters. To select the optimal number of hidden layer units, this paper compares the MAPE values under different hidden units. First, select

the number of hidden layer units in [5, 10, 15, 30, 50] for testing, and their MAPE values are 0.8300%, 0.3818%, 0.5834%, 0.4834% and 0.3850%, respectively. Therefore, the number of hidden layer units (10) with the smallest MAPE value (0.3818%) is selected as the number of hidden layer units in this experiment. For different time steps, the STA-FNet model proposed in this paper has different evaluation performances. To select the best time step for the next experiment, this model tests the RMSE, MAE, and MAPE under the time step [5, 10, 15, 20, 25, 30], and the corresponding evaluation distribution is shown in Fig. 5. It can be seen that the RMSE, MAE and MAPE values of the model are relatively minimal when the time step is 20. Therefore, the time step selected for this experiment is 20.

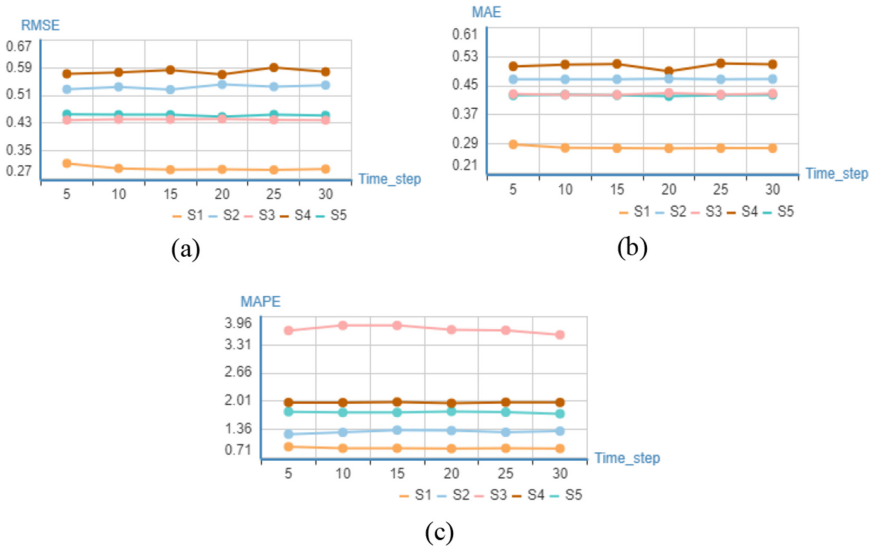

Fig. 5. Evaluation values at different time steps.

Performance Analysis. After determining the parameters of the model, use the designed training set and test set to verify the prediction performance of the model. The loss function curve generated by the model in 5 different road segments during the training process is shown in Fig. 6. It is not difficult to find that with the increase in the number of iterations, the loss function curves of the training set and the test set decrease rapidly and gradually converge, which indicates that the design of the model is reasonable.

Comparison of Baseline Methods. To further compare the prediction performance of different algorithms, CS-LSTM, DCS-LSTM, and BiLSTM are selected for comparison with the STA-FNet model. It can be seen from Fig. 7 and Table 1 that the proposed model outperforms the baseline models after 200 iterations. The average MAPE of the 5 road segments is 2.7320%, and the prediction accuracy is 97.2620%

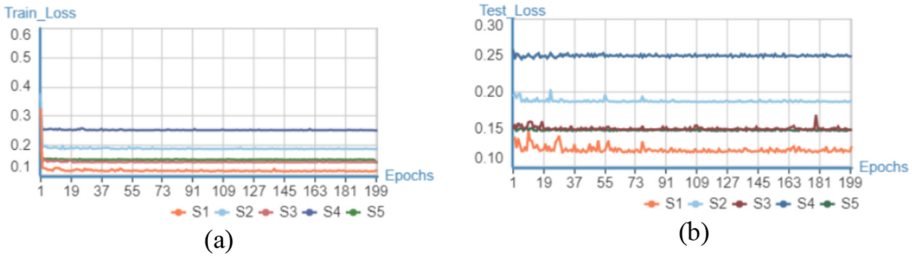

Fig. 6. Training set loss and Test set loss for 5 road segments with Epochs of 200.

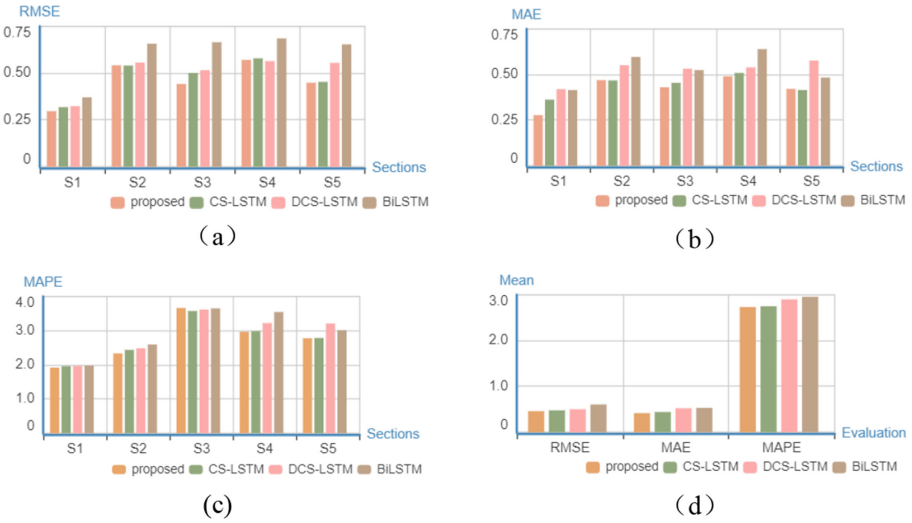

Fig. 7. RMSE(a), MAE(b), MAPE(c) and Mean(d) of Different Model Predictions.

The average MAPE values of STA-FNet, DCS-LSTM, and CS-LSTM are 0.2235%, 0.2065%, and 0.0563% higher than those of BiLSTM, respectively. The experimental results show that modeling the position interaction of surrounding vehicles and exploring the interaction between surrounding vehicles and target vehicles have a positive effect on improving the accuracy of vehicle trajectory prediction. In addition, DCS-LSTM, which introduces a convolutional social pooling layer, achieves the same prediction effect as CS-LSTM, while maintaining more accurate vehicle local position information, and its prediction performance is further improved.

The average MAPE value of the algorithm proposed in this paper is 0.0170% higher than that of DCS-LSTM. Because the algorithm in this paper takes into account the historical trajectory of the target vehicle itself and the historical location interaction between the target vehicle and the surrounding vehicles, while further considering the current traffic congestion of the road. At the same time, the accuracy of the prediction is further improved by using the Spatio-temporal attention mechanism acting on the whole process of extracting temporal and spatial features of the data. The comprehensive

Table 1. RMSE, MAE, and MAPE of Different Model Predictions.

Models	Evaluation	S1	S2	S3	S4	S5	Mean
STA-FNet (proposed)	RMSE	0.2968	0.5420	0.4425	0.5705	0.4487	**0.4601**
	MAE	0.2781	0.4694	0.4305	0.4901	0.4217	**0.4180**
	MAPE	1.9198	2.3349	3.6612	2.9660	2.7783	**2.7320**
DCS-LSTM	RMSE	0.3185	0.5405	0.501	0.5792	0.4535	0.4785
	MAE	0.3628	0.4679	0.4547	0.5092	0.4153	0.4420
	MAPE	1.9628	2.4374	3.5738	2.985	2.7858	2.7490
CS-LSTM	RMSE	0.3234	0.5565	0.5157	0.564	0.555	0.5029
	MAE	0.4201	0.551	0.5323	0.5393	0.5768	0.5240
	MAPE	1.9667	2.4839	3.6167	3.2207	3.2081	2.8992
BiLSTM	RMSE	0.3708	0.657	0.6651	0.6854	0.6536	0.6064
	MAE	0.4153	0.5967	0.5243	0.6395	0.4839	0.5319
	MAPE	1.9764	2.5963	3.6505	3.5435	3.0106	2.9555

consideration and analysis of the influencing factors around the target vehicle further enhance the interpretability of vehicle trajectory prediction.

In summary, the STA-FNet model proposed in this paper outperforms other models in predicting vehicle trajectories. The model starts from the historical trajectory of the target vehicle retrospectively and also explores the influence of the spatial distribution of neighboring vehicles on the future decision of the target vehicle. In addition, the impact of the road congestion state on the trajectory of the target vehicle at the current moment is quantified. The above influencing factors are combined to study the future trajectory of the target vehicle, which makes the model more comprehensive and interpretable in terms of trajectory prediction.

6 Conclusion

To effectively predict the motion trajectories of target vehicles in complex traffic scenes, this paper proposes a Fusion Neural network with the Spatio-Temporal Attention (STA-FNet) model. The novelty of this model is that it comprehensively considers the effects of attention factors, Spatio-temporal feature relationships among data, and the decision-making effects of the interaction between the target vehicle and surrounding vehicles on vehicle trajectory prediction. The comparative experimental results show that the prediction accuracy of the STA-FNet model constructed in this paper is 97.2620%, which is 0.0170%, 0.1672%, and 0.2235% higher than the DCS-LSTM model, CS-LSTM model, and BiLSTM model respectively. The effect is significantly better than other models, and the model has a stronger interpretability interpretation. In addition, objectively speaking, this paper sacrifices the complexity of the model to improve prediction accuracy.

The next step will consider reducing the complexity of the model based on improving or maintaining the existing prediction accuracy, which will be considered in the next research. Long-term prediction of vehicle trajectories enables a more accurate prediction of vehicle trajectories.

References

1. Liang, Y., Zhao, Z.: Vehicle trajectory prediction in city-scale road networks using a direction-based sequence-to-sequence model with spatiotemporal attention mechanisms. arXiv e-prints, arXiv: 2106.11175 (2021)
2. Leon, F., Gavrilescu, M.: A review of tracking and trajectory prediction methods for autonomous driving. Mathematics **9**(6), 660 (2021)
3. Huang, Z., Wang, J., Pi, L., et al.: LSTM based trajectory prediction model for cyclist utilizing multiple interactions with environment. Patt. Recogn. **112**, 107800 (2021)
4. Wang, J., Wang, P., Zhang, C., Su, K., Li, J.: F-Net: fusion neural network for vehicle trajectory prediction in autonomous driving. In: 2021 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 4095–4099 (2021)
5. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 2015 Proceedings of 3rd *International Conference* on *Learning* Representations (ICLR), pp. 1049–0473 (2015)
6. Cai, Y., Wang, Z., Wang, H., et al.: Environment-attention network for vehicle trajectory prediction. IEEE Trans. Veh. Technol. **70**(11), 11216–11227 (2021)
7. Yu, D., Lee, H., Kim, T., et al.: Vehicle trajectory prediction with lane stream attention-based LSTMs and road geometry linearization. Sensors **21**(23), 8152 (2021)
8. Wang, S., Shao, C., Zhai, Y., et al.: A Multifeatures Spatial-Temporal-Based Neural Network Model for Truck Flow Prediction. J. Adv. Transp. **2021** (2021)
9. Lin, L., Li, W., Bi, H., et al.: Vehicle trajectory prediction using LSTMs with spatial-temporal attention mechanisms. IEEE Intell. Transp. Syst. Mag. **14**(2), 197–208 (2021)
10. Xi, P., Gu, Y.: Research on expressway travel time prediction based on deep learning. In: 2021 Fifth International Conference on Traffic Engineering and Transportation System (ICTETS), SPIE, 12058, pp. 399–403 (2021)
11. Jin, J., Guo, H., Xu, J., et al.: An end-to-end recommendation system for urban traffic controls and management under a parallel learning framework. IEEE Trans. Intell. Transp. Syst. **22**(3), 1616–1626 (2020)
12. Deo, N., Trivedi, M.M.: Convolutional social pooling for vehicle trajectory prediction. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops (ICCVPR), pp. 1468–1476 (2018)
13. Zhang, H., Wang, Y., Liu, J., et al.: A multi-modal states based vehicle descriptor and dilated convolutional social pooling for vehicle trajectory prediction. arXiv preprint arXiv:2003.03480 (2020)
14. Li, T., Ni, A., Zhang, C., et al.: Short-term traffic congestion prediction with Conv–BiLSTM considering spatio-temporal features. IET Intel. Transport Syst. **14**(14), 1978–1986 (2020)