DeepPTP: A Deep Pedestrian Trajectory Prediction Model for Traffic Intersection

Zhiqiang Lv[1], Jianbo Li[1*], Chuanhao Dong[1], Yue Wang[1], Haoran Li[1], and Zhihao Xu[1]
[1] College of Computer Science & Technology, Qingdao University
Shandong, 266071, China
[e-mail: lijianbo@qdu.edu.cn]
[*]Corresponding author: Jianbo Li

Received January 21, 2021; revised May 26, 2021; accepted July 1, 2021; published July 31, 2021

Abstract

Compared with vehicle trajectories, pedestrian trajectories have stronger degrees of freedom and complexity, which poses a higher challenge to trajectory prediction tasks. This paper designs a mode to divide the trajectory of pedestrians at a traffic intersection, which converts the trajectory regression problem into a trajectory classification problem. This paper builds a deep model for pedestrian trajectory prediction at intersections for the task of pedestrian short-term trajectory prediction. The model calculates the spatial correlation and temporal dependence of the trajectory. More importantly, it captures the interactive features among pedestrians through the Attention mechanism. In order to improve the training speed, the model is composed of pure convolutional networks. This design overcomes the single-step calculation mode of the traditional recurrent neural network. The experiment uses Vulnerable Road Users trajectory dataset for related modeling and evaluation work. Compared with the existing models of pedestrian trajectory prediction, the model proposed in this paper has advantages in terms of evaluation indicators, training speed and the number of model parameters.

Keywords: Pedestrian trajectory, traffic intersection, interactive features, pure convolutional network

The code is open sourced in https://github.com/qdu318/DeepPTP.
This research was supported in part by National Key Research and Development Plan Key Special Projects under Grant No. 2018YFB2100303, Shandong Province colleges and universities youth innovation technology plan innovation team project under Grant No. 2020KJN011, Shandong Provincial Natural Science Foundation under Grant No. ZR2020MF060, Program for Innovative Postdoctoral Talents in Shandong Province under Grant No. 40618030001, National Natural Science Foundation of China under Grant No. 61802216, and Postdoctoral Science Foundation of China under Grant No.2018M642613.

http://doi.org/10.3837/tiis.2021.07.002 **ISSN : 1976-7277**

1. Introduction

Pedestrian trajectory prediction is an important research field of smart city. It is beneficial to urban travel planning, alleviation of traffic congestion and urban commercial planning. Pedestrian trajectories are mainly divided into long-term trajectory in Location-Based Social Networks (LBSN) [1] and short-term trajectory for continuous location. LBSN is a network for social interaction established through smart terminal devices. It is an online platform for pedestrians to share information about interests, hobbies, status and activities. LBSN provides pedestrians with location-based services, which allows pedestrians to share their location details in social networks. Combining with auxiliary factors such as interest and hobbies, the task of long-term trajectory prediction of pedestrian is based on pedestrian sign-in data to mine the potential patterns and laws, and then predict the possible situations of pedestrian locations. The research of short-term trajectory prediction mainly focuses on the calculation of a series of spatial and temporal characteristics of pedestrians' continuous position changes in a short time. The trajectory of vehicles on the road is restricted by road distribution and traffic signs. The regularity of the trajectory route can make the model more convenient to calculate the relationship. However, pedestrian trajectories are affected by surrounding pedestrians and nearby static obstacles in real life, which are complex, changeable and less restricted by traffic rules. The above process poses harder challenges to pedestrian short-term trajectory prediction.

The intersection of a road is the intersection of two or more roads. It is the place where vehicles and pedestrians meet, turn and evacuate. Traffic control designs (such as signal lights and other traffic management facilities) are often installed at intersections to ensure traffic safety and smooth flow. The mining and prediction of pedestrian trajectory data in road intersections is conducive to rational organization and planning of traffic at intersections. More importantly, it is also an indispensable link in improving traffic capacity and ensuring traffic safety. Trajectory prediction can be regarded as a sequence generation task, which predicts future trajectories based on past positions. Traditional recurrent neural network cannot handle the problems of exponential explosion and vanishing gradient in the recursive process, and it is difficult to capture long-term time correlation. However, combining different LSTMs can solve this problem well. Recurrent neural network can describe the behaviors of dynamic time. Unlike feedforward neural network that accepts more specific structure input, recurrent neural network cyclically transmits state in its own network. So, it can accept a wider range of time series input, as shown in **Fig. 1**. However, the recurrent neural network has a fatal problem that each element in the sequence is directly related to all the elements in front of the current element and its complexity will explode in the calculation process, as shown in (1).

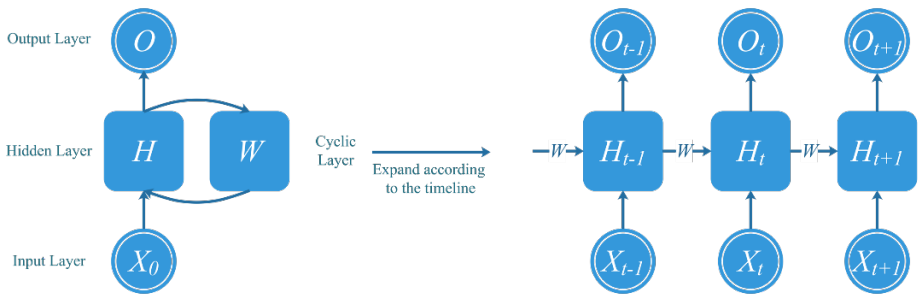

Fig. 1. Basic process of recurrent neural network. *X* is the input of the network. *O* is the output of the network. *H* is the value of hidden layer. *W* is the weight maintained by the unit.

$$\exists P(X) = \prod_{i=1}^{N} P\left(x_i | x_{1,\ldots,x_{i-1}}\right) \quad (1)$$

This paper proposes a deep pedestrian trajectory prediction model (DeepPTP) for the task of pedestrian short-term trajectory prediction. It uses the continuous position data to predict the trajectory direction of pedestrians at intersections. DeepPTP uses Spatial Convolution Layer to calculate the correlation of spatial feature and uses Temporal Convolution Layer to calculate the dependence of temporal feature. The main contributions of this paper are as follows:

- This paper proposes using pedestrian trajectory data to predict the direction of pedestrian trajectory at intersections. The model design method transforms the existing data regression mode into a mode that integrates data regression and classification.
- This paper proposes a deep spatial-temporal model, which combines trajectory data, distance difference, time difference, interaction among pedestrians and other auxiliary factors to realize multi-modal data prediction tasks.
- This paper uses the Vulnerable Road Users (VRU) trajectory dataset for experimental evaluation. The results of experimental evaluation show that DeepPTP has higher accuracy and training speed in the task of short-term pedestrian trajectory direction prediction. Compared with the existing model of pedestrian trajectory prediction, the accuracy of model maximum increases by 29.86%.

2. Related Works

The main task of pedestrian trajectory prediction is to build a model of time data to capture the long-term dependence of the trajectory. Compared with the process of vehicles driving in the road lanes, the short-term trajectory of pedestrian is freer and more complex. The motion state of pedestrian has lesser impact on the restrictions of traffic signs and rules. So, the short-term trajectory prediction task of pedestrians is more challenging [2]. Interaction among pedestrians is an important aspect that affects trajectory. The Attention Mechanism Based Pedestrian Trajectory Prediction Generation Model (AttenGAN) [3] builds pedestrian interaction mode and makes probabilistic multi-mode prediction. It uses the encoding-decoding structure to predict the possibility of pedestrian short-term trajectory. The special attention mechanism calculates the state of other pedestrians which is an implicit factor that affects the main features. The experimental results prove that AttenGAN has the capabilities of comprehensive pedestrian interaction prediction and can predict the future trajectory of joint and multiple possibilities. The work [4] divides the pedestrian trajectory into a grid and calculates the fuzzy spatial membership of each point in the trajectory to generate fuzzy sub-trajectories. The fuzzy sub-trajectories are input into a model composed of two sets of Fuzzy-LSTM to calculate the proximity and periodicity of the trajectory. The work [5] uses a bidirectional long-term and short-term recurrent neural network to calculate the hidden state during pedestrian movement and uses a dual attention module to calculate individual movement information and group interaction information of pedestrians that have a greater impact on the trajectory. The work [6] proposes a recurrent neural network (Social-Scene-LSTM, SS-LSTM) with three levels of scale. In order to increase the accuracy of the prediction, SS-LSTM uses a circular shape neighborhood setting instead of the traditional rectangular shape neighborhood. It fully simulates the interaction among groups in social activities. Kothari et al. [7] proposed a knowledge-based data method (TrajNet++) for large-scale interaction between pedestrian trajectories. The TrajNet++ is composed of motion encoding-

decoding module and interactive module. The motion coding-decoding module is responsible for feature calculation of pedestrian historical information. The interaction module is used to capture the interaction state between pedestrians.

The trajectory prediction of pedestrians at intersections can be summarized as the distribution prediction of a person's possible multiple paths in the future. [8][9][10] This kind of research focuses on how to use the existing historical data to choose the future direction. Liang et al. [11] propose a probabilistic model (Multiverse) to calculate the future trajectory of pedestrians based on historical locations and scenes. Firstly, the Multiverse displays the position on a multi-scale grid to obtain a multi-modal representation of the future position. Secondly, it predicts the offset of each grid unit through the fine scale of features. The CoverNet [8] uses the current and past states of the moving target to calculate the multi-modal probability distribution of the future state. It limits the possible future state set of the moving target within a reasonable prediction range, so as to achieve the maximum possible choice of the target's trajectory at the intersection.

The DeepPTP draws on the calculation model of spatial correlation and temporal dependence of the above research. It integrates a variety of auxiliary factors and interaction characteristics among pedestrians to increase the temporal and spatial characteristics of the data. In the end, we build a probability model for the future trajectories of pedestrians at intersections. which opens up a new direction for the tasks of pedestrian short-term trajectory prediction.

3. Model Design

3.1 Definition

Target Problem

The target problem of this work is to use pedestrian trajectory data at intersections to predict the final trajectory direction of pedestrians. The model of traditional pedestrian trajectory prediction uses a collection of historical locations of pedestrians to predict future location information. The above process is a kind of regression problem research. However, due to the complexity of pedestrian movement, the accuracy rate and the applicability of the model are not optimistic. This work transforms the pedestrian trajectory prediction into a classification problem from a regression problem. Taking a typical intersection with four directions as an example, we assign digital signs (*0*, *1*, *2*, *3*) to the four directions. The dividing line in the four directions is the red line, as shown in the **Fig. 2**. According to the time distribution of the pedestrian trajectory, the trajectory is marked with corresponding numbers. The final position of the trajectory is the main basis for the marking process. For example, the blue trajectory is marked as *0*, the yellow trajectory is marked as *3*, and the green trajectory is marked as *2*. Taking the center of the intersection as the origin of the coordinate system, we map the road to a coordinate system with the center lines of the four roads as the X and Y axes. The position nodes of each trajectory are mapped to coordinate system data. If the final position of the trajectory is within the four red dividing lines, it will be regarded as noise data.

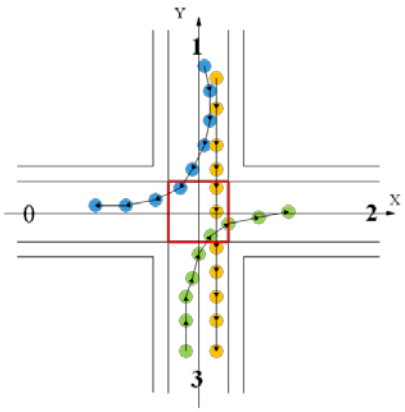

Fig. 2. Definition of intersection direction. The nodes of different colors are the track positions of pedestrians. The number is the direction sign given to the intersection.

Design of Input and Output Data

The data for this work is the trajectory data of each pedestrian, which is composed of multiple location points. In order to be able to calculate the trajectory data of multiple pedestrians at the same time, we design the input data into the following basic shape: [*bs*, *pn*, *c*]. The *bs* represents the batch size of each epoch and it also represents the number of pedestrian trajectories participating in the calculation at the same time. The *pn* represents the number of points that make up the pedestrian trajectory. Since the total distance of each pedestrian trajectory and the number of location points are inconsistent, we need to preprocess the *pn* of each trajectory. The main idea of preprocessing is to use a zero matrix to supplement the difference between the current trajectory and the trajectory with the most position points. The *c* represents the channels of the data, which contains two sequences ($\{X\}$ and $\{Y\}$ of pedestrian trajectory). The final output data is designed into the following basic shape: [*bs*, *r*]. The output data contains the final trajectory directions of multiple pedestrians. The *r* represents the trajectory direction predicted by the model, which is the number marked in **Fig. 2** (e.g., 0, 1, 2, 3)

3.2 Main Structure

The main structure of the model is shown in the **Fig. 3**. Firstly, the sequence data of pedestrians at the intersection is input into the Spatial Convolution Layer to enhance the spatial correlation of the primary data features in the intersection. Secondly, the reciprocal relationship between the auxiliary factor and the pedestrian is fused with the primary data features for multi-dimensional embedding [20]. Finally, the Temporal Convolution Layer captures the temporal dependence of the data to form high-dimensional features. In order to reduce the computational complexity of the model, the internals of the main modules that make up DeepPTP are all based on convolutional neural networks, including Spatial Convolution Layer, Temporal Convolution Layer and Full Connected Network.

Two sequences ($\{X\}$ and $\{Y\}$) form a trajectory sequence and they respectively represent the coordinates of the track position point. The trajectory sequence is standardized before the model training process. The Z-Score standardization method is suitable for situations where the maximum and minimum values of the data are unknown, or there are outliers that exceed the value range. Different from the driving mode of vehicles, human behaviors are complex and changeable. These situations that pedestrians travel on a curve or find road shortcuts

happen from time to time. Therefore, Z-Score is more suitable for the application scenarios of pedestrian trajectory prediction. During the experiment, we also tried other standardization methods [14], such as Min-Max and Decimal Scaling. But the results after using these standardization methods are not ideal.

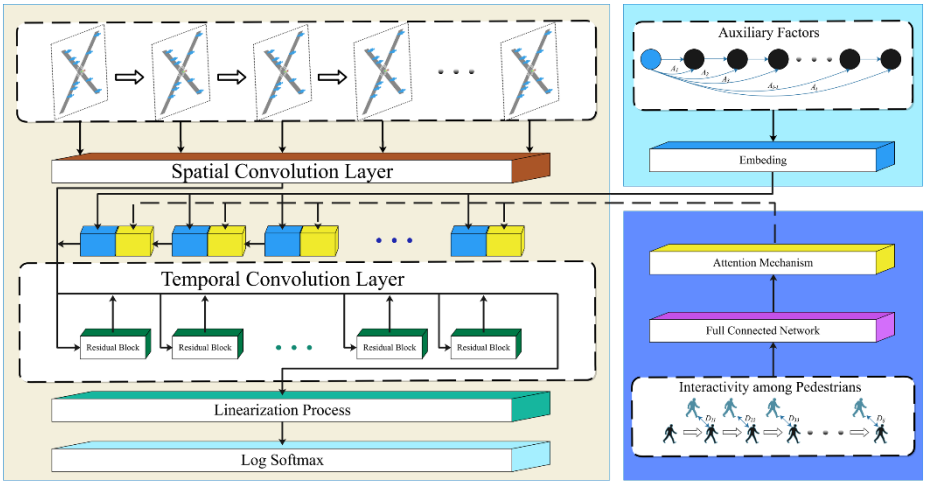

Fig. 3. Basic process of DeepPTP.

3.3 Spatial Convolution Layer

The main function of spatial convolution calculates the spatial correlation of the trajectory sequence. The shallow network of spatial convolution extracts low-level features. As the network deepens, low-level features are fused to form high-level features, and the spatial invariance of features can be maintained. The data feature contains two sequences: $\{X\}$ and $\{Y\}$. These two sequences are concatenate as a data matrix. Linearization stands for data linearization processing, which can change the data matrix to the input dimension of Conv1d. Local connections enable Conv1d to extract local features of the data. The process of convolution is to do template matching in each local area of the matrix. The pooling operation of Conv1d is a down-sampling process. The relative relationship among different features plays an important role. It can increase the generalization ability of the model by calculating the translation relationship and controlling overfitting.

3.4 Auxiliary Factors and Interactivity among Pedestrians

The auxiliary factor is calculated based on the differences between the first node and other nodes in the trajectory sequence, as shown in (2). A represents the sequence of different auxiliary factors. We use two auxiliary factors: distance and time. The distance sequence increases the spatial correlation among nodes, which enables the model to indirectly obtain the magnitude of pedestrian trajectory changes. Since the trajectory of the pedestrian is mapped to the coordinate system, the distance relationship between the nodes can be directly calculated by the distance equation between two points in the coordinate system, as shown in (3). The addition of the time sequence to the feature calculation process makes the model no longer consider one-sidedly the spatial relationship. Time sequence increases the time dependence among nodes. It balances the weight ratio of trajectory features in the spatial and temporal relationship.

$$A_i = a_i - a_0, i \in [0, len(a)] \tag{2}$$

$$d_i = \sqrt{(x_i - x_0)^2 + (y_i - y_0)^2} \tag{3}$$

Observing the traffic rules and signal light restrictions, pedestrians usually cross the road together in the course of the intersection. Therefore, the movement behavior of an individual pedestrian is not only related to his own position information, but also related to the individual pedestrians around him. In order to capture the influence of surrounding pedestrians on a certain pedestrian, we use the Attention mechanism to select the location information of other pedestrians that have an impact on the current pedestrian, as shown in (4) and (5).

$$D_{ij}^t = \{x_j^t - x_i^t, y_j^t - y_i^t\} \tag{4}$$

$$O = Attention\langle FC(D_{ij}^t; W_{fc})\rangle \tag{5}$$

The relative position (D_{ij}^t) of a person and the current pedestrian is mapped to the feature space from the spatial coordinate system by the Fully Connected Network (*FC*). The W_{fc} is the parameter of the Fully Connected Network. The Attention mechanism selects the characteristics of other individuals that have an impact on the current pedestrian by strengthening the correlation characteristics and reducing the non-correlated characteristics among the trajectories.

3.5 Temporal Convolution Layer

The role of the temporal convolution layer enhances the temporal features of pedestrian trajectories by means of complex winding machines. The temporal convolution layer uses the idea of large-scale parallel processing of convolutional neural networks, which maps multi-dimensional matrices to time series. It obtains a large enough receptive field through a multilayer network and performs deep network parallel processing. The model needs to be based on sufficiently long-term data to establish temporal features, especially for pedestrian trajectories. The pedestrian trajectory of this work is composed of position points that change in accordance with time. Longer-term data means more location points for each trajectory. In a multi-layer network, the important purpose of this work to set different ranges of receptive fields is to preserve the influence of all historical locations on the final features.

The temporal convolution layer is composed of multiple residual blocks. The residual block is connected through a shortcut connection to make the data easier to optimize. If a nonlinear unit $f(x, \Theta)$ is used to approximate the objective function $h(x)$, the objective function can be split into the identity function x and the residual function $h(x) - x$. According to the general approximation theorem, a non-linear unit composed of a neural network has enough ability to approximate the original objective function. Therefore, the original problem is transformed into that the linear unit $f(x, \Theta)$ approximates the residual function $h(x)$ - x and the $f(x, \Theta)$ + x approximates the $h(x)$. The multi-layer residual structure has been proved to be beneficial to the nonlinear process of deep neural networks, while the performance of the single-layer residual structure is not ideal. The residual block mainly includes two processes that are causal convolution and dilation convolution. The **Fig. 4** shows the causal convolution and dilation convolution [12] in the Temporal Convolution Layer.

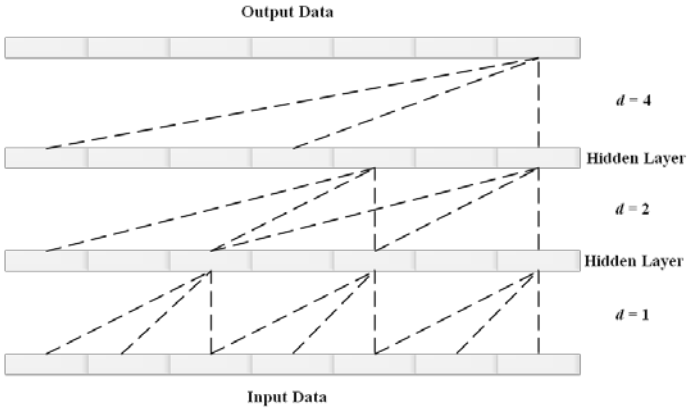

Fig. 4. Basic process of Temporal Convolution Layer. The *d* is the dilation factor. The dotted line is the data transfer process.

We can know that the value of neuron node of each layer only depends on the historical value of corresponding node of previous layer, which reflects the idea of causal convolution. The neuron nodes in each layer represent the temporal features of pedestrian position nodes. The feature of the future location node can only rely on the data information of the pedestrian before the node, which is in line with the normal derivation idea. The equation of causal convolution is shown in (6). $\{x_1, x_2, ..., x_t\}$ is the input sequence. $\{y_1, y_2, ..., y_t\}$ is the hidden layer output sequence. $\{f_1, f_2, ..., f_k\}$ is the filter sequence. Causal convolution only pays attention to historical information and ignores future information. The result of y_t is derived from the data before x_t. The larger the K is, the more historical information can be traced back. If the original input sequence of the current layer is [0, i], the original input sequence of the next layer will become [0, i+1].

$$y_t = \sum_{i=1}^{k} f_i \cdot x_{t-k+i} \tag{6}$$

The dilation convolution is used to expand the range of the receptive field. The layer-by-layer dilation factor d increases exponentially by 2. The extraction of the information of the previous layer is skipped in each layer, which reflects the idea of dilation convolution. We expect to capture the multi-scale contextual information of the historical trajectory, which means that the current location node can contain the features of different historical nodes to varying degrees. However, this kind of influence among nodes will gradually decrease or even disappear are transferred in the recursive unit in the traditional recurrent neural network as the features. The equation of the dilation convolution is shown in (7). The d is the dilation factor, which changes by an exponent of *2* according to the depth of the network. Increasing d or K can increase the range of the receptive field. The result of each layer at time t can only be calculated by the data at time [0, t], which reflects the idea of causal convolution. The result at each moment is obtained by jumping to the value of the dilation factors in the previous layer of the network, which reflects the idea of dilation convolution.

$$y_t = \sum_{i=0}^{k-1} f_i \cdot x_{t-d\times i} \tag{7}$$

3.6 Normalization Function

The *Softmax* function performs a series of index calculations. When the input data of *Softmax* is large, the feature calculation will produce overflow phenomenon. Similarly, when the input data is negative and its absolute value is large, the numerator and denominator of *Softmax* will become very small and the data will be approximately 0. The above process will cause underflow phenomenon. The experimental data of this work are continuous location points. For each trajectory, the coordinates of the location points that compose it usually change continuously. Therefore, the features of part of the endpoints of the trajectory often have the phenomenon of overflow or underflow due to excessive or small weights. Compared with the traditional normalization function *Softmax*, The *Log Softmax* adds a *Log* operation on the basis of *Softmax*. It can not only solve the problems of data overflow and underflow, but also improve the speed and stability of calculation process. The proof process of the above theory is as shown in equation (8). M is the maximum value of $\{x_i\}$. The maximum value of *exp* is 0 after subtracting M from any x_i, so the *Log Softmax* is unlikely to have an overflow phenomenon. At the same time, at least one of the summations in the final expression is 1, so this prevents the expression from underflow. *Log Softmax* converts the complex exponential calculations of *Softmax* into addition and subtraction calculations to improve calculation efficiency. The position data of the pedestrian trajectory in this experiment are floating point numbers in the coordinate system. The result of these data through *exp* operation has more invalid 0 after the decimal poin, which affects the validity of the data. The use of *log* method effectively removes invalid 0 placeholders and improves the validity of numbers

$$\begin{aligned} log(Softmax) &= log\left(\frac{exp(x_i)}{\sum_j^n exp(x_j)}\right) \\ &= log\left(\frac{\frac{exp(x_i)}{exp(M)}}{\frac{exp(x_1)}{exp(M)}+\frac{exp(x_2)}{exp(M)}+\cdots+\frac{exp(n)}{exp(M)}}\right) \\ &= log\left(\frac{exp(x_i-M)}{\sum_j^n exp(x_j-M)}\right) \\ &= log\big(exp(x_i-M)\big)-log\big(\sum_j^n exp\big(x_j-M\big)\big) \\ &= (x_i-M)-log\big(\sum_j^n exp\big(x_j-M\big)\big) \end{aligned} \tag{8}$$

4. Experiment

4.1 Data Preparation

In this experiment, we use the VRU trajectory dataset [16][17][18]. The dataset is recorded by cameras and LiDAR [19] at urban intersections. It contains 1068 pedestrian trajectories and 464 cyclist trajectories. The true distribution of the trajectories is shown in the **Fig. 5**. The trajectory of pedestrian is recorded by a wide-angle stereo camera system. The trajectory of the cyclist is recorded by tracking the center of the cyclist through LiDAR. The complete dataset contains 1532 files in CSV format. Each file contains a VRU trace. In order to fit our research direction, we remove the trajectory data whose trajectory endpoint is located in the center of the intersection.

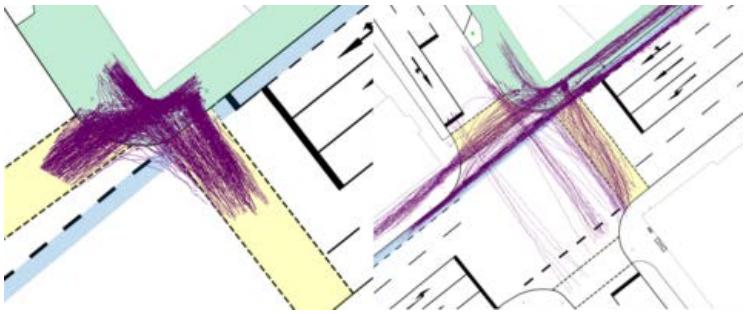

Fig. 5. The distribution of pedestrian trajectories on the actual map. The figure on the left is the trajectory of pedestrian. The figure on the right is the trajectory of the bicycle.

4.2 Performance Comparison

Evaluation Indicator

The confusion matrix is a situation analysis table for the prediction results of the classification model in deep learning. It summarizes the data of dataset in a matrix according to the real category and the category predicted by the model. The rows of the confusion matrix represent the true values. The columns of the confusion matrix represent the predicted values, as shown in the **Fig. 6**. In this experiment, we use the confusion matrix for multi-classification tasks. The positive state represents the accurate classification result of the current prediction process. The negative state represents the sum of the error classification results of the current prediction process.

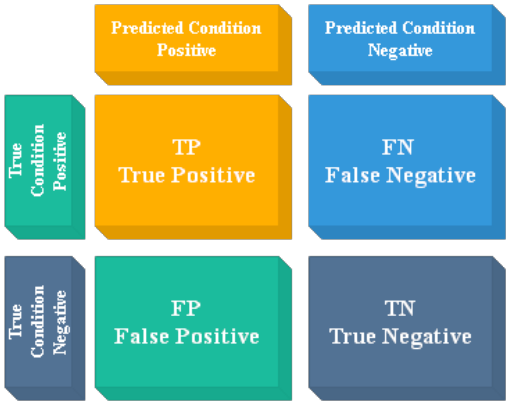

Fig. 6. Confusion matrix calculation process. *TP* is the number of positive classes predicted by the model as positive classes. *FN* is the number of positive classes predicted by the model as negative classes. *FP* is the number of negative classes predicted by the model as positive classes. *TN* is the number of negative classes predicted by the model as negative classes.

In this experiment, we use four evaluation indicators to compare the performance of DeepPTP and other existing models. The *Accuracy* is an evaluation index for traditional classification problems. It represents the percentage of the total sample that the model predicts correctly. However, the *Accuracy* has an obvious drawback. When the data is very uneven in categories, *Accuracy* cannot objectively evaluate the pros and cons of the model. Therefore, we additionally add the *Precision*, *Recall* and *F1 Score* as evaluation indicators. The *Precision* represents the probability of the sample that is actually positive among all the samples that are

predicted to be positive. *Recall* represents the probability of being predicted to be a positive sample in a sample that is actually positive. The *Precision* and *Recall* indicators [15] sometimes trade each other. The higher the *Precision* is, the lower the *Recall* is. The *Precision* and *Recall* must be balanced in some scenarios. The most common method is *F1 Score* in view of the above situation. The calculation methods of the four evaluation indicators are shown in the **Table 1**.

Table 1. Calculation methods of evaluation indicators

Indicators	Calculation Methods
Accuracy	(TP + TN) / (TP + TN + FP + FN)
Precision	TP / (TP + FP)
Recall	TP / (TP + FN)
F1 Score	(2 × Precision × Recall) / (Precision + Recall)

Ablation Experiment

In order to verify the role of each layer of DeepPTP, we design an ablation experiment and show the experimental results in **Fig. 7**. We use linear transformation to replace each layer of ablation module to ensure the consistency of model dimensions in the modeling process. The linear transformation changes the dimensionality of the data to make the dimensionality of the data suitable for the input requirements of the Temporal Convolution Layer. The Temporal Convolution Layer plays a major role in DeepPTP, which establishes a temporal dependent correlation with the trajectory of pedestrians. The DeepPTP without Temporal Convolution Layer has the worst prediction accuracy, and its final *F1 Score* is only maintained at around 0.65. The final *F1 Score* of DeepPTP without Spatial Convolution Layer is maintained at around 0.75. This experimental result shows the predictive ability of DeepPTP without calculating the spatial correlation of the data. In this work, the trajectories of pedestrians are distributed at the intersection of the road. This spatial distribution always exists, because the movement process of pedestrians has always been restricted by the distribution of roads. Therefore, the Spatial Convolution Layer has played an active role after being added to DeepPTP. Although the role of the auxiliary factor is not very obvious, it can also help to improve the prediction accuracy. Especially in the early stage of training, the auxiliary factor can make the model converge quickly and obtain a certain accuracy rate. This is because the auxiliary factor retains the initial characteristics of the pedestrian trajectory and it gives the model low-level features to form multi-scale information fusion with the high-level feature of model.

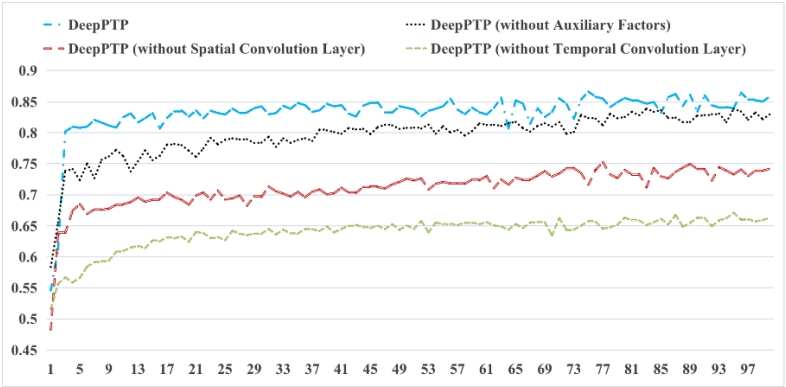

Fig. 7. Results of ablation experiments. The abscissa is the number of training epochs. The ordinate is the value of *F1 Score*.

Baseline Comparison

We set up some baselines to verify the performance of DeepPTP, as shown in the **Table 2**. In addition to 3 popular recurrent neural networks, we have implemented 6 special models for pedestrian prediction. We can know that DeepPTP has the highest accuracy compared to the existing models. It achieves the best results among the four evaluation indicators. Compared with the traditional recurrent neural network (RNN, LSTM and GRU), the evaluation indicator maximum increases by 84.08%. The comparison result of DeepPTP and recurrent neural network proves that the combination of spatial correlation and temporal dependence can significantly improve the performance of the model compared to the mode of only temporal dependence. Compared with the special models (Fuzzy-LSTM, Encoder-Decoder, AttenGAN, TrajNet++, CoverNet and Multiverse) ranking, the evaluation indicator maximum increases by 29.86%. The comparison between DeepPTP and other models with special structures proves the performance advantage of the combination of convolutional neural network in trajectory prediction. DeepPTP-3 and DeepPTP-5 indicate that the Temporal Convolution Layer of DeepPTP has 3 and 5 hidden layers, respectively. In DeepPTP, increasing the number of layers of Temporal Convolution Layer can improve the accuracy of the model. However, the increase in the number of layers sacrifices the speed of training. In actual application scenarios, the ratio of model training speed and accuracy should be considered comprehensively.

Table 2. Result of evaluation indicators

Models	Evaluation Indicators			
	Accuracy	Precision	Recall	F1 Score
RNN	0.4775	0.4712	0.4857	0.4783
LSTM	0.4912	0.4896	0.4961	0.4928
GRU	0.4836	0.4802	0.4973	0.4886
Encoder-Decoder	0.4928	0.4912	0.5032	0.4971
SS-LSTM	0.5289	0.5156	0.5273	0.5214
Fuzzy-LSTM	0.6769	0.6874	0.6946	0.6910
AttenGAN	0.6978	0.6865	0.6722	0.6793
TrajNet++	0.7345	0.7249	0.7347	0.7298
CoverNet	0.7641	0.7844	0.7743	0.7793
Multiverse	0.7977	0.7964	0.8076	0.8020
DeepPTP-3	0.8597	0.8514	0.8476	0.8317
DeepPTP-5	**0.8790**	**0.8667**	**0.8712**	**0.8689**

Analysis and Expansion

Although the traditional recurrent neural network can calculate and derive the weight of time series data, its calculation process and parameters are single. Using complex and special Deep Learning models can help researchers freely control the participation of model parameters to achieve the effect of improving prediction accuracy. The Encoder-Decoder and AttenGAN structure can solve the problem of different lengths of pedestrian trajectory points. They can encode trajectories of different lengths into trajectory features of the same length, which avoiding the problem of data sparseness caused by the original zero supplement operation. The length of the sequences encoded by the Encoder-Decoder and AttenGAN are not dynamic, so the long trajectory has information loss in this experiment. The Fuzzy-LSTM emphasizes the

periodicity of pedestrian movement, but we cannot set the periodic characteristic of pedestrians in the choice of road intersection trajectory. The SS-LSTM, TrajNet++, CoverNet, and Multiverse have their own calculation methods for pedestrian interaction characteristics. However, these calculation methods are too complicated and the applicable scenarios are too fixed. We found that too many auxiliary calculations make the weight ratio of the main location features of the data too low in the weight of the entire model when we migrate the intersection data set to these trajectory models, which makes the prediction accuracy of the model poor. The DeepPTP uses a completely convolutional structure to avoid the high computational complexity of traditional recurrent neural networks and it separates pedestrian interaction characteristics from other auxiliary factors to increase the proportion of pedestrian position and pedestrian interaction in the overall model weight.

The **Fig. 8** shows the actual distribution of the three confusion matrices. From the experimental results, we can know that the three models have better prediction results for the fourth category. This is because the fourth category has more data in the actual data set than other categories and its sample features are rich. Although there is less data in the first category, DeepPTP shows good accuracy. This is because DeepPTP makes full use of the historical information of the trajectory and all historical information participates in the calculation process of the Temporal Convolution Layer. This process fully excavates the position information from the starting point of the pedestrian to the center of the intersection.

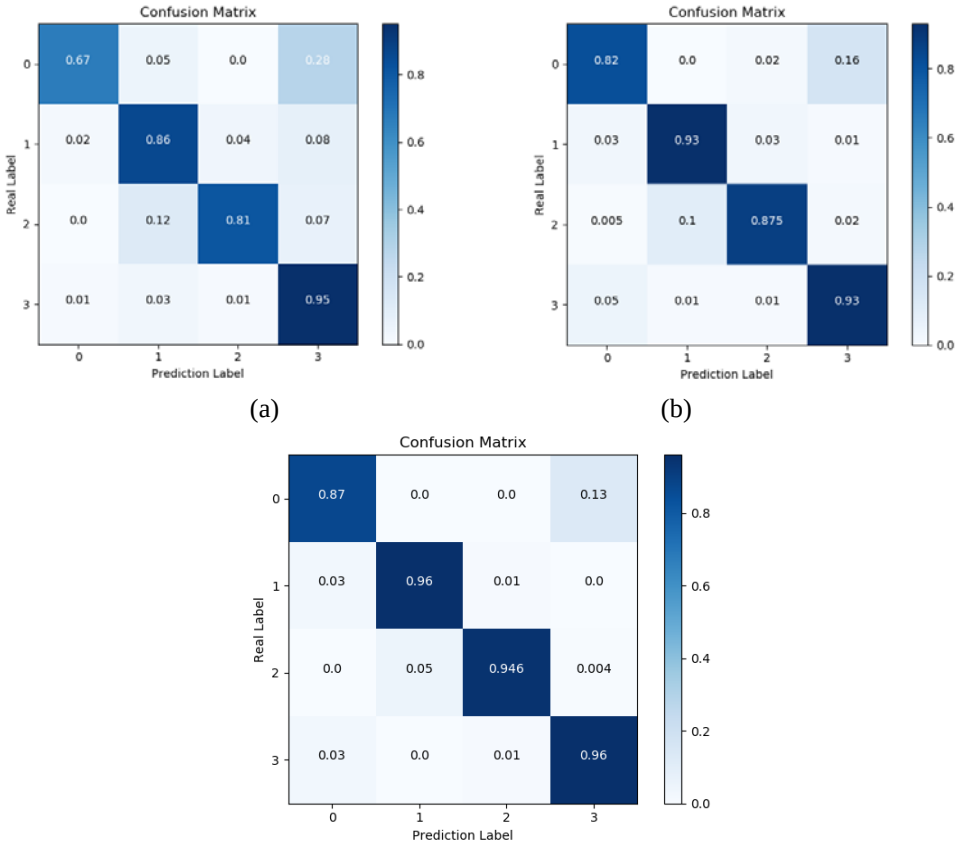

Fig. 8. Distribution of prediction results in confusion matrix. The (a) represents AttenGAN. The (b) represents Multiverse. The (c) represents DeepPTP.

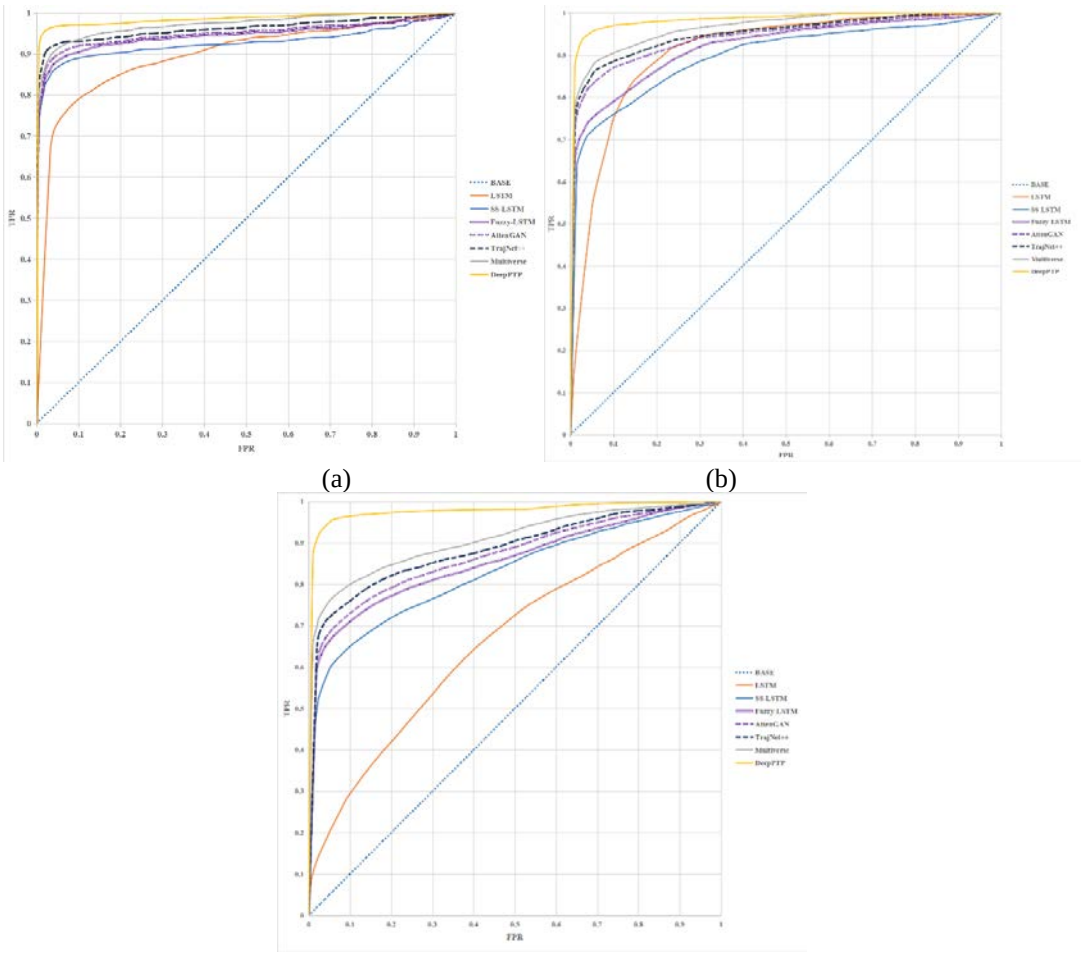

Fig. 9. Distribution of prediction results in the ROC curve. (a) is the result of two classifications, (b) is the result of three classifications and (c) is the result of four classifications.

The curve of Receiver Operating Characteristic (ROC) [13] shows the accuracy of model intuitively. The closer the ROC curve is to the upper left corner, the higher the recall rate of the model, which represents the points on the ROC curve have the least classification errors. In order to compare the performance of different models, we draw the curve of ROC of each model to the same coordinate system, as shown in the **Fig. 9**. The (a) is the prediction result of two types of trajectories and the directions of the two types of trajectories are left turn and right turn respectively. The three types of trajectory directions of (b) are turn left, turn right and go straight. The (c) includes the four directions of the intersection. The DeepPTP has obvious accuracy advantages. It is worth mentioning that as the number of classification categories increases, the accuracy of the above models has decreased to varying degrees. However, DeepPTP has always maintained high accuracy. The structure of the spatial-temporal model makes it retain spatial correlation and temporal dependence. The DeepPTP uses dynamically changing receptive fields and longer-term historical information, so it can more accurately capture the spatial-temporal relationships and recursive relationships among the various nodes of the pedestrian trajectory.

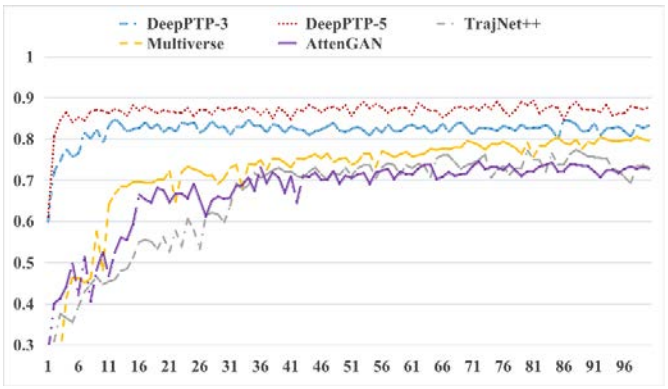

Fig. 10. Changes in F1 Score between DeepPTP and current models. The abscissa is the number of training epochs. The ordinate is the value of *F1 Score*.

As mentioned earlier, the dilation convolution mechanism and pure convolution structure of model strongly improve the model convergence speed of the training process. Most of the existing special models are designed based on traditional recurrent neural networks, which cannot avoid the single-step calculation mode. The *F1 Score* changes of the above models and DeepPTP in training process are shown in the **Fig. 10**. We can know that the performance of Multiverse, AtteenGAN and TrajNet++ in the early stage of model training changes slowly. The performance of the models reaches the optimum at the end of the entire training process. The powerful model convergence ability of DeepPTP allows the model to reach the optimal performance value in the early stage of training. During the entire model training process, the model doesn't appear to over-fit phenomena. The comparison of the experimental results of DeepPTP-3 and DeePTP-5 proves that the increase of the number of Temporal Convolution Layers can improve performance within a certain range and doesn't affect the high robustness of DeepPTP.

Deep learning relies on numerous network parameters in the neural network to participate in the calculation. It has the disadvantages of complex network structure, large amount of calculation, slow speed and it is difficult to transplant to embedded devices. As the network model has deeper and deeper layers and more and more parameters, reducing model size and computational loss is crucial. The causal convolution mechanism of DeepPTP realizes the tracing and fusion of historical information at the feature level. Unlike the models mentioned above, which only calculate historical information at the data level, this process results in models having huge numbers of parameters. The DeepPTP is a lightweight fully convolutional model. Compared with other special pedestrian prediction models, The DeepPTP has the smallest amount of model parameters, as shown in the **Fig. 11**.

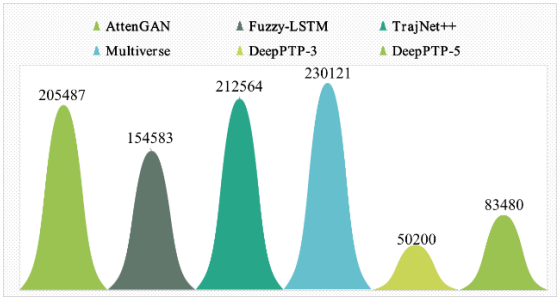

Fig. 11. Comparison of the number of model parameters between DeepPTP and existing models. The abscissa is the name of model. The ordinate is the number of parameters for each model.

5. Conclusion

This paper proposes a prediction model of deep pedestrian trajectory. The model uses pedestrian trajectories at traffic intersections to achieve the short-term prediction of pedestrian trajectory direction. We use the Temporal Convolution Layer instead of the traditional recurrent neural network to overcome the single-step calculation mode of the traditional recurrent neural network. The causal convolution and dilation convolution mechanisms of Temporal Convolution Layer not only improve the training speed, but also enable the model to trace longer historical information. The combination of multiple auxiliary factors and trajectory position information further strengthens the spatial correlation and temporal dependence, and improves the accuracy of the model. In future work, we will focus on the impact of pedestrian and vehicle trajectory directions on traffic in different directions at intersections. The prediction of long-term and short-term traffic in different directions at intersections helps to realize the intelligence of traffic lights and the automation of traffic command. The trajectories of pedestrians and vehicles at intersections have similar characteristics. Research on multi-modal models that can be applied to both pedestrian and vehicle traffic prediction can take advantage of the complementarity among multiple modalities, eliminate the redundancy among modalities and learn better feature representations.

References

[1] C. Su, S. Peng, X. Xie, and N. Liu, "Study on Check-in Prediction Based on Deep Learning and Factorization Machine," *Computer Science*, vol. 46, no. 5, pp. 185-190, May. 2019. Article (CrossRef Link)

[2] M. Goldhammer, S. Köhler, S. Zernetsch, K. Doll, B. Sick, and K. Dietmayer, "Intentions of Vulnerable Road Users − Detection and Forecasting by Means of Machine Learning," *IEEE Trans. on Intelligent Transportation Systems*, vol. 21, no. 7, pp. 3035-3045, Jul. 2020. Article (CrossRef Link)

[3] Y. Sun, Q. Jiang, J. Hu, J. Qi, and Y. Peng, "Attention Mechanism Based Pedestrian Trajectory Prediction Generation Model," *Journal of Computer Applications*, vol. 39, no. 3, pp. 1-9, Sept. 2018. Article (CrossRef Link)

[4] M. Li, H. Zhang, P. Qiu, S. Chen, J. Chen, and F. Lu, "Predicting Future Locations with Deep Fuzzy-LSTM Network," *Acta Geodaetica et Cartographica Sinica*, vol. 47, no. 12, pp. 1660-1669, Dec. 2018. Article (CrossRef Link)

[5] Y. Ou, Q. Shi, X. Wang, and L. Wang, "Pedestrian Trajectory Prediction Based on GAN and Attention Mechanism," *Laser & Optoelectronics Progress*, vol. 57, no. 14, pp. 1-12, Jul. 2020. Article (CrossRef Link)

[6] H. Xue, D. Q. Huynh, and M. Reynolds, "SS-LSTM: A Hierarchical LSTM Model for Pedestrian Trajectory Prediction," in *Proc. of the IEEE Winter Conference on Applications of Computer Vision*, Lake Tahoe, NV, USA, pp. 1186-1194, 2018. Article (CrossRef Link)

[7] P. Kothari, S. Kreiss, and A. Alahi, "Human Trajectory Forecasting in Crowds: A Deep Learning Perspective," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1-15, Apr. 2021. Article (CrossRef Link)

[8] T. Phan-Minh, E. C. Grigore, F. A. Boulton, O. Beijbom, and E. M. Wolff, "Covernet: Multimodal Behavior Prediction using Trajectory Sets," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 14062-14071, 2020. Article (CrossRef Link)

[9] B. Völz, H. Mielenz, I. Gilitschenski, R. Siegwart and J. Nieto, "Inferring Pedestrian Motions at Urban Crosswalks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 2, pp. 544-555, Feb. 2019. Article (CrossRef Link)

[10] D. Ridel, E. Rehder, M. Lauer, C. Stiller, and D. Wolf, "A Literature Review on the Prediction of Pedestrian Behavior in Urban Scenarios," in *Proc. of the 21st International Conference on Intelligent Transportation Systems*, Maui, HI, USA, pp. 3105-3112, 2018. Article (CrossRef Link)
[11] J. Liang, L. Jiang, K. Murphy, T. Yu, and A. Hauptmann, "The Garden of Forking Paths: Towards Multi-Future Trajectory Prediction," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 10505-10515, 2020. Article (CrossRef Link)
[12] S. Bai, J. Z. Kolter, and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," *arXiv*, pp. 1-14, Apr. 2018. Article (CrossRef Link)
[13] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach," *Biometrics*, vol. 44, no. 3, pp. 837-845, Sept. 1988. Article (CrossRef Link)
[14] S. Helama, M. Lindholm, M. Timonen and M. Eronen, "Detection of climate signal in dendrochronological data analysis: a comparison of tree-ring standardization methods," *Theoretical and Applied Climatology*, vol. 79, no. 3, pp. 239-254, Aug. 2004. Article (CrossRef Link)
[15] M. Buckland and F. Gey, "The relationship between recall and precision," *Journal of the American society for information science*, vol. 45, no. 1, pp. 12-19, Jan. 1994. Article (CrossRef Link)
[16] M. Goldhammer, E. Strigel, D. Meissner, U. Brunsmann, K. Doll and K. Dietmayer, "Cooperative multi sensor network for traffic safety applications at intersections," in *Proc. of 15th International IEEE Conference on Intelligent Transportation Systems*, Anchorage, AK, USA, pp. 1178-1183, 2012. Article (CrossRef Link)
[17] M. Bieshaar, G. Reitberger, S. Zernetsch, B. Sick, E. Fuchs and K. Doll, "Detecting Intentions of Vulnerable Road Users Based on Collective Intelligence," *arXiv*, pp. 1-20, Sep. 2018. Article (CrossRef Link)
[18] M. Bieshaar, S. Zernetsch, A. Hubert, B. Sick, and K. Doll, "Cooperative Starting Movement Detection of Cyclists Using Convolutional Neural Networks and a Boosted Stacking Ensemble," *IEEE Transactions on Intelligent Vehicles*, vol. 3, no. 4, pp. 534-544, Dec. 2018. Article (CrossRef Link)
[19] F. Y. Lin and J. M. Liu, "Chaotic lidar," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 10, no. 5, pp. 991-997, Dec. 2004. Article (CrossRef Link)
[20] G. Vladuca and A. Tudora, "Prompt fission neutron spectrum calculations for n+ 238U reaction using the multi-modal model," *Annals of Nuclear energy*, vol. 28, no. 16, pp. 1643-1652, Nov. 2001. Article (CrossRef Link)

Zhiqiang Lv was born in Weifang city, Shandong province, China in 1995. He received the bachelor degree in software engineering from Ludong University, China, in 2019. From 2016 to 2019, he studied in the software parallel group of the State Key Laboratory of Computer Architecture, Institute of Computer Technology Chinese Academy of Sciences. He is studying for the master degree in Qingdao University, China, majoring in computer technology. He has won 13 awards of the science and technology innovation in university, 1 invention patent and many university student scholarships. His main research directions are traffic demand research based on traffic trajectory and travel time, traffic data forecast research based on traffic flow, speed and traffic congestion, high performance parallel computing research based on deep learning and reinforcement learning.

Jianbo Li was born in Weifang city, Shandong province, China in 1980. He received the Ph.D. degree in computer science and technology department from the University of Science and Technology of China in 2009. From 2013 to 2014, he was a visiting scholar at Fordham University. He is currently the professor of the college of computer science & technology in Qingdao University and the director of the Institute of ubiquitous network and urban computing of the Qingdao University. He is the chairman of ACM Qingdao Branch, deputy secretary general of Qingdao Computer Society, senior member of China Computer Federation, and member of Internet of Things Professional Committee of China Computer Federation. His research interests include urban computing, mobile social networks and data offloading.

Chuanhao Dong was born in Jining city, Shandong province, China in 1993. He received the bachelor degree in electronic commerce in Taiyuan University of Science and Technology, China, in 2019. From 2015 to 2019. Now, He is studying for the master degree in Qingdao University, China, majoring in computer technology. His main research directions are traffic flow predictions and congestion prediction based on historical traffic information and geospatial information using deep learning methods.

Yue Wang was born in Zibo city, Shandong province, China in 1997. She received the bachelor's degree in computer science and technology from Harbin Normal University in June 2020. Now, she is studying for the master's degree in Qingdao University, China, majoring in computer technology. She has won many university student scholarships. Her research focuses on using deep learning to predict traffic flow among regions.

Haoran Li was born in Heze city, Shandong Province in 1992. He is studying for the master degree in Qingdao University, China, majoring in computer technology. His research focuses on urban computing, urban area classification and traffic big data mining.

Zhihao Xu was born in Binzhou city, Shandong province, China, in 1992. He received the bachelor degree in digital media technology from Qingdao University, China, in 2014. He is studying for the master degree in Qingdao University, China, majoring in computer technology. His main research directions are deep learning, urban computing, Intelligent Transportation, crowd density research and crowd interest prediction.