Multi-PPTP: Multiple Probabilistic Pedestrian Trajectory Prediction in the complex junction scene

Linhui Li, Bin Zhou, Jing Lian, *Member, IEEE*, Xuecheng Wang, Yafu Zhou

Abstract—**Pedestrian trajectory prediction in dynamic, multi-agent traffic junction scene is an important problem in the context of self-driving cars. Accurately predicting the trajectory of the agent, especially the pedestrian with high mobility and randomness, is of great significance to autonomous driving technology. In this paper, we propose Multi-PPTP, a novel trajectory prediction model that first utilizes a composite raster map to model the complex represen tations and interactions of road components, including dynamic obstacles(e.g., pedestrians, vehicles, and cyclists) and static road context information(e.g., lanes, traffic lights, and sidewalks), then uses MapNet and AgentNet to extract spatio-temporal features by deep convolutional networks and LSTM to automatically derive relevant features, and next constructs Interaction-AttNet aggregate features to learn the high interactions among all components by affine transformation and multi-head attention mechanism. Additionally, we also propose a series of unique loss functions to predict multiple possible trajectories of each pedestrian while estimating their probabilities. Following extensive offline evaluation and comparison to the state-of-the-art baselines, our approach outperforms the state-of-the-art on our large scale in-house dataset significantly.**

Index Terms—**Interaction, trajectory prediction, autonomous driving, traffic junction scene.**

I. Introduction

THE data analysis of pedestrian behavior is of great significance for road safety, robot navigation, safety monitoring and other fields [1]. Research on pedestrian trajectories requires collecting pedestrian data for offline analysis to understand pedestrian behavior and surroundings, then make reasonable decisions. In a system with real-time decision-making capabilities, correct prediction of the pedestrian's path can issue alarms and take corresponding preventive measures as soon as possible.

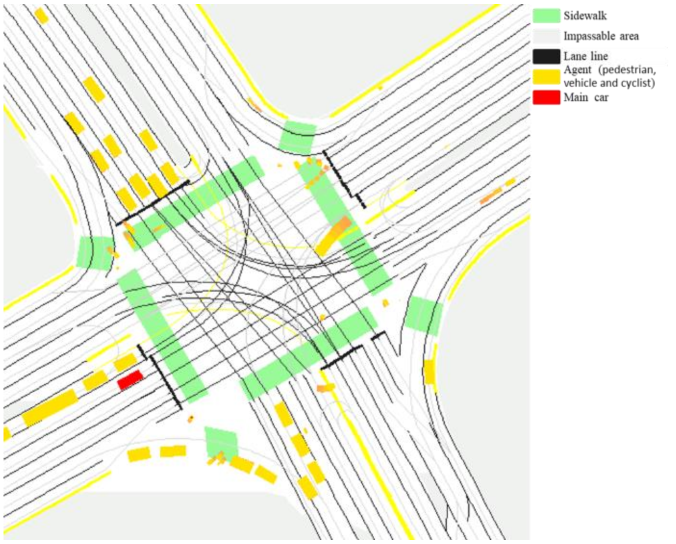


Fig. 1. Illustration of the rasterized rendering to represent high-definition map and agent trajectories. In this map, different RGB values represent different semantic information, where green is for sidewalk, grey is for impassable area, black is for lane line, yellow is for agents in this scene where the rectangles of different sizes represent different types of agents and red is for main car.

In this paper, we focus on human behavior prediction in complex junction scene, which is a crucial task for self-driving vehicles in real-world environments. The junction scene involves complex interactions of various types of agents (pedestrians, vehicles, and cyclists), different junction types (T-shape or X-shape junction), and mutation road conditions. The trajectory prediction problem can be regarded as a sequence decision problem, which is predicting the trajectory of the pedestrian in the future according to the position of pedestrian in the past. But this problem is very complicated:

1) The behavior of the pedestrian has a certain degree of randomness, because future state prediction is inherently stochastic, as human can not know each other motivations. Generating a deterministic trajectory in the prediction task is not in line with the actual situation. Therefore, it is very important to consider multiple results and possibilities.

2) Each pedestrian does not exit independently. According to the research of Moussaid *et al.* [2], 70% of pedestrians tend to walk in groups, and they interact in the same space and time with other agents and surroundings, which makes it more difficult to deal with this problem, and junction scene has the most complex interactions among vehicle, pedestrian, cyclist and surroundings.

We seek a model that can handle interactions in complex scenarios and predict multiple outcomes and their likelihoods. High Definition maps (HD-maps) can provide useful semantic and geometric information for predicting human behavior, as

This work was supported by the National Natural Science Foundation of China (GrantNos.51775082, 61976039, 61473057) and the China Fundamental Research Funds for the Central Universities (Grant Nos. DUT19LAB36 DUT17LAB11), and Science and Technology Innovation Fund of Dalian (2018J12GX061).

Linhui Li, Bin Zhou, Jing Lian, Xuecheng Wang, and Yafu Zhou are with the Dalian University of Technology, Dalian 116086, China. (e-mail: {lilinhui, dlut_zhoubin, lianjing, wangxuecheng, dlzyf}@dlut.edu.cn).

Linhui Li, Jing Lian, and Yafu Zhou are also with the Key Laboratory of Energy Conservation and New Energy Vehicle Power Control and Vehicle Technology.

Corresponding author: Jing Lian (e-mail: lianjing@dlut.edu.cn)

the behaviors of the human largely depend on the map topology. For example, pedestrians rarely pay attention to cars coming from another lane, so effective use of HD-maps is necessary for trajectory prediction. In our work, we take the specially processed HD-maps as the prior knowledge of the model. Recent works use deep learning to learn semantic representations from maps. Most of these approaches are using original image or feature map. In our work, the HD-maps is not only highly structured, but also combines static and dynamic semantic information: road semantics (e.g., lane connection lines, stop lines, available areas, etc.), traffic light information and other agents in the past trajectory information (Fig. 1). See the Sec3 for specific design details.

Chai *et al*. [3] used cluster analysis of the trajectory to capture the intent uncertainty and control uncertainty of the trajectory. It may be useful for agents that are more constrained by road scenes and traffic rules. But it is unrealistic for pedestrians to follow a specific pattern of behavior. So, we propose a multi-modal prediction branch outputs the final motion forecasting. For each pedestrian, we predict K possible future trajectories and their confidence scores.

In summary, our contributions are:

1) We propose a complete multiple probability trajectory prediction model for pedestrians at complex traffic scene which directly demonstrates how to build a complete composite raster map to integrate scene context information and agent dynamic information for trajectory prediction.

2) We present a novel Interaction-AttNet (Fig. 2) to integrate interaction features and design unique loss function for the multiple probability prediction task.

3) We evaluate the proposed method on our in-house behavior prediction dataset, and we demonstrate empirically that our model effectively improves the prediction accuracy while improving the prediction speed, and achieves the effect of reducing model redundancy and improve model interpretability.

II. Related work

In this section, we review works on trajectory prediction. The existing pedestrian trajectory prediction methods can be divided into two categories: one is a model-based method, and the other is a data-driven deep learning method.

Schneider *et al*. [4] compared the method based on a single kinematic model with the method based on multi-model interaction. In [5], it provides a non-parametric model for probability prediction by assuming that hidden variables obey Gaussian distribution. Kitani *et al*. [6] proved that the semantic information of the static environment (e.g., the extended area of the sidewalk and the position of the zebra crossing, etc.) helps the model to predict the pedestrian future trajectory accurately, and they treated this problem as Markov Decision Process and learned one-step policy. Yamaguchi *et al*. [7] used an optimization-based method to manually design an objective function covering motion-related aspects to optimize the function parameters in real time. However, such model-based methods are limited in two aspects when dealing with trajectory prediction problems:

1) It has to manually design the model function to simulate the scene of human interaction, instead of fitting the function through data, which makes the model function only suitable for simple interaction scenes (e.g., gravity/repulsion social force model);

2) Model focuses on establishing the interaction between pedestrians at present, but cannot reasonably model and predict the interaction behaviors that will occur in the future.

In recent years, with the rapid development of computer vision and artificial intelligence, the pedestrian trajectory prediction method based on deep learning has attracted the attention of researchers. Based on the Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) [8] methods have been proved to be effective in dealing with sequence problems, but the method based on RNN and LSTM has a great disadvantage is that model cannot effectively model

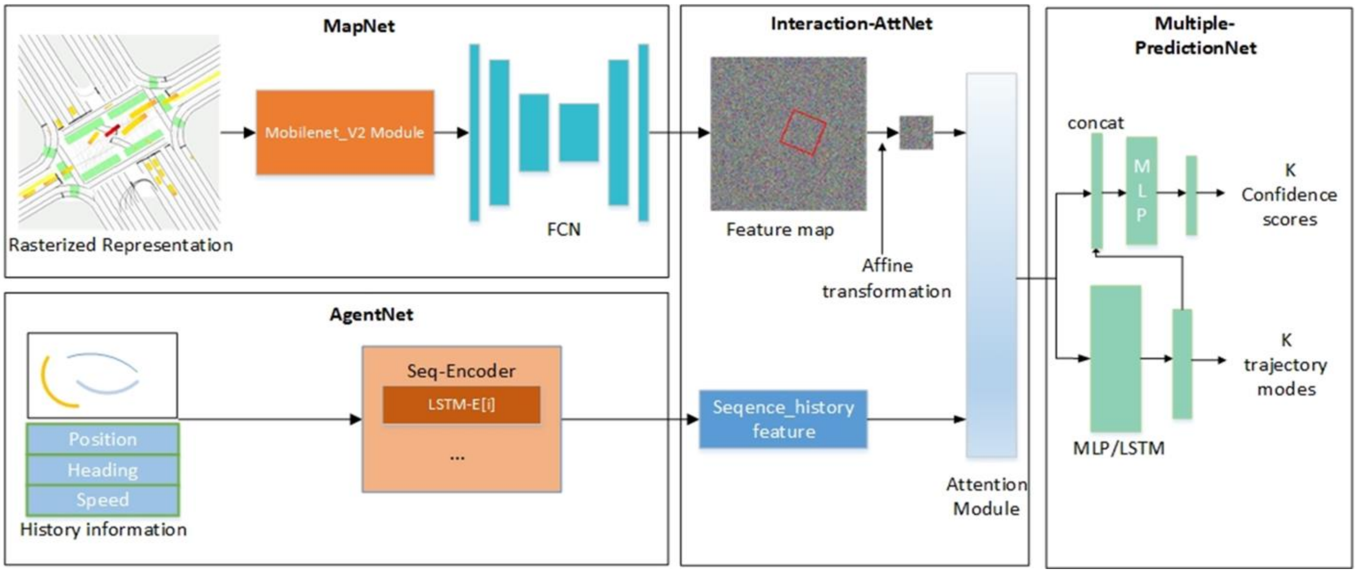


Fig. 2. A review of our proposed model. We construct rasterized representation (Fig. 1) maps and use history sequence information to obtain topology, semantic and history sequence feature by MapNet and AgentNet. Such features are then passed to Interaction-AttNet to model the high-order interactions by using affine transformation and multi-head attention mechanism. We compute two types of branches to predict future trajectories and their scores.

the spatial connection between pedestrians. In [9], the pedestrians are modeled as a space-time graph, where edges (time and space) are connected to RNNs. Time edges capture the information of a single pedestrian and space edges capture the information of pedestrian interaction. The output adopts a bivariate Gaussian distribution. This method can effectively model spatio-temporal information, but the calculation is complicated and expensive. Alahi *et al*. [10] proposed social LSTM model. By meshing the surrounding pedestrians, the different characteristics of the surrounding pedestrians are calculated. Gupta *et al*. [11] introduced a method based on Generative Adversarial Network (GAN) [12] into the field of pedestrian trajectory prediction, and used the maximum pooling method to generate multiple socially acceptable trajectories, but the features extracted by this method are the largest feature after the pooling model, the model ignores other useful feature information for pedestrian interaction. By fusing the contextual information of the scene in the environment and the historical trajectory of pedestrians, Sadeghian *et al*. [13] used the GAN network to generate multiple acceptable trajectories under physical conditions. Vineet *et al*. [14] proposed the graph attention network (GAT) [15], [28] into the field of trajectory prediction, and enhanced the reasoning ability of trajectory prediction by using the GAT. In [16] work, they modeled the pedestrian trajectories as a spatio-temporal graph to replace the aggregation layers and used temporal CNNs [17] and graph neural networks (GNN) [18] to break through the limitations of recursive architecture, such as LSTM.

Similar to our work, a few of previous work [19], [20], [30] directly models the probability distribution to predict multiple trajectories. But these works cannot make good use of known information, and they also use simple fusion or concatenate to interact with agent behaviors and intentions, and this is not representative. These works focus on predicting vehicle trajectories which is difference with pedestrian prediction problems, in contrast, our model is designed for pedestrian problems. Cheng *et al*. [21] used a conditional variational autoencoder (CVAE) to predict the diversity of trajectories, simplified the distribution of noise to a Gaussian distribution by introducing variational estimation, and used the posteriori distribution after assuming the distribution as the prediction model, and then trained the prediction model. However, current network models based on data-driven methods have some problems such as increased network redundancy, decreased interpretability, low real-time performance, and insufficient accuracy.

III. Problem Statement and Model Overview

A. Notation and problem formulation

We express the trajectory prediction problem as estimating the state of the pedestrian in the future based on the state and information of the pedestrian in the past. In the following, at any time instant *t*, the i^{th} person in the scene is represented by his/her xy-coordinates (x_i^t, y_i^t). We use indices $i, j \in \{1, \ldots, N\}$ to refer to pedestrians, where N is the total number of pedestrians in scenes. We assume that the trajectory of each agent is influenced by the prior movements, the location of other traffic agents and the physical constraints, as well as the traffic rules. So, for the junction scene, the inputs of our model are twofold: scene information I and trajectory information X_i, encoded by 4 dimensional vectors, $X_i = \{(x_i^t, y_i^t, v_i^t, \psi_i^t) \in R^2 | t = 1, \ldots, t_{obs}\}$. Given these input features and the true trajectory of each pedestrian $Y_i = \{(x_i^t, y_i^t) \in R^2 | t = t_{obs+1}, \ldots, t_{pre}\}$, our model will predict the trajectory of each pedestrian $\hat{Y}_i = \left\{\left(\widehat{x_i^t}, \widehat{y_i^t}\right) \in R^2 \middle| t = t_{obs} + 1, \ldots, t_{pred}\right\}$ in the future by learning the parameters W′ of model.

$$\widehat{Y_\iota} = f(X_i, I; W') \tag{1}$$

Among them, the model parameter W′ is the weight of all deep neural networks used in the model. Use backpropagation algorithm to train all weight parameters of the model, and achieve stochastic gradient descent by minimizing the loss function $\mathcal{L}$ for pedestrian predicted trajectory and real trajectory.

In the next section, we will introduce our approach. We first describe how to characterize scene and traffic agent trajectories information in composite rasterized map. Next, we use pedestrian attribute information as input of the AgentNet in parallel (e.g., position, speed, and heading angle). And then we combine these features by unique interaction attention model (Interaction-AttNet). We also design Multiple-PredictionNet and present multitask loss function to predict multiple probability trajectories.

B. Constructing composite raster maps

In our work, we adopt a composite raster map to represent the HD-map information and agent trajectories information. As we all know, when we walk on the street, we are always influenced by surrounding physical environments (e.g., lane line, traffic light and obstacle information) and agent in traffic environments (e.g., pedestrian, vehicle, and cyclist). In this paper, the composite raster map is composed of static semantic map, dynamic agent location map and real-time traffic light map.

As shown in Fig. 3, static map includes passable and impassable areas, lane line, sidewalk and stop line information, which are represented by different RGB values. The dynamic agent location map is a map sequence of obstacle historical location, every time step has one channel. And the traffic light map includes a real-time traffic light information. These elements make up our composite raster map. This simple map format not only provides basic geometric and semantic features, but also innovatively integrates time series information for accurate forecasting. This is also in line with the direct feelings of pedestrians making decisions.

C. MapNet: exacting spatio-temporal representation

We present a novel network, called MapNet, to learn the complex topological relationship from the composite raster map. As shown in Fig. 2, MapNet has two modules: first, we use Mobile_net_V2 to extract a feature representation for the whole scene *I* and then use FCN module to further extracts the topological information of the map, and return a feature map V_p as same size as the input image.

$$V_p = MapNet(I; W_p) \tag{2}$$

Refer to (2), W represent the weights of the model. An important benefit of using such branch model is that the network can fully extract spatio-temporal representation.

D.AgentNet: extract traffic participant representation

AgentNet receives the past pedestrians state X_i as input, and uses LSTM to extract their sequence feature. For each pedestrian, we first embed the input feature into a higher dimension using a multilayer perceptron (MLP), and then using LSTM to encode these pedestrian states generating vector $V_s(i)$ for pedestrian i.

$$V_s(i) = LSTM(MLP(X_i; W_{MLP}), h_{LSTM}(i); W_{LSTM}) \tag{3}$$

E. Interation-AttNet

In this section, we propose a novel network to combine the extracted information from AgentNet and MapNet. As mentioned above, the behaviour of pedestrian depends on surroundings and interaction with other obstacles and whole map. In the previous jobs [10], [11], [22], [23], they used either permutation invariant symmetric functions, such as max, mean or gumbel pooling functions, or ordering functions such as sorting based on Euclidean distance [13]. By using these methods, models can deal with the interaction with other agents well. But these methods either need to discard some features which may be very important for interaction, or need to set the maximum number of pedestrians to maintain the consistency of the data dimension, or introduce a priori bias, such as using Euclidean distance to measure the strength of interactivity [16].

The attention mechanism [24] is derived from the human visual attention mechanism. Humans use limited attention resources to extract effective information from a large amount of information, which is a long-term survival mechanism that humans have evolved. The attention mechanism based on deep learning was first applied to the image field [25]. The mechanism is similar to the human visual attention mechanism. It obtains the most effective and critical information for the task goal from a large number of information. Afterwards, the attention mechanism has been widely used in natural language processing fields such as speech recognition and machine translation. With the help of the Encoder-Decoder model [26], good results have been achieved in related fields [27]. The attention module in the article is similar to the human visual attention mechanism, which will pay more attention to the information of interest. When observing an image, you do not understand all of the image at once, but first focus on the local features of the image. Similar to observing images, when dealing with complex scenes of human-human interaction, pedestrians usually combine the current state and the influence of the surrounding pedestrians, and focus on the part of the information that affects them, so as to quickly make corresponding decisions and change trajectories [2].

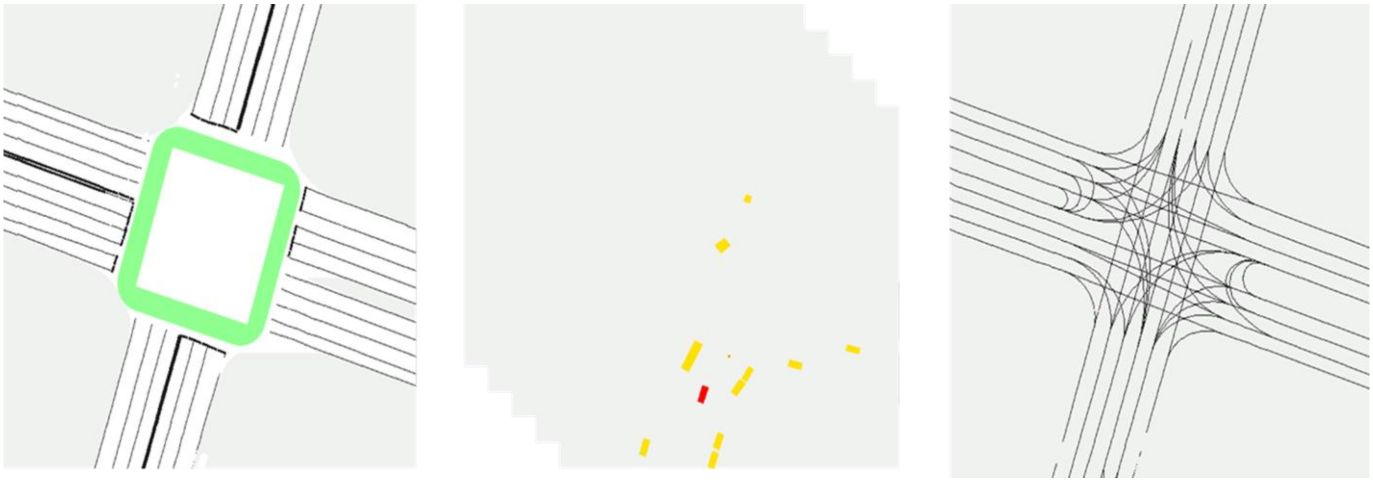
To avoid these flaws, in this paper, we have innovatively proposed a method to handle the interaction between pedestrians and the environment. As shown in Fig. 2, the Interaction-AttNet receives AgentNet's output $V_i(s)$ and MapNet's output $\mathrm{V_p}$, whole model consists of two architectures that first one extracts the pedestrian-centric features V_p' through affine transformation from V_p. And then using muti-head attention network [27] to combine the affine feature V_p' and the sequence feature $V_i(s)$.

$$V_p' = Affine(V_p) \tag{4}$$

$$C_p(i) = ATT\left(V_p', V_s(i); W_{ATT}\right) = \frac{1}{M}\sum\nolimits_m c_p(i) \tag{5}$$

$$c_p(i) = softmax\left(\frac{V_p' \cdot MLP(V_S(i))}{\sqrt{d_{V_p'}}}\right) \cdot V_p' \tag{6}$$

Where $ATT(\cdot)$ is multi-head attention mechanism [27] which consists of multiple scaled dot product attention modules



(a) HD static map (b) Historical location of agents (c) Traffic light illustration

Fig. 3. The composition of raster map. Illustration of the rasterized rendering to represent high-definition map and agent trajectories, where (a) is HD static map which contains the topology and semantic information in the junction scene, (b) is historical location of different agents which contains 5 channels represents 5 frames and (c) is the dynamic traffic light graph that represents traffic information of different lanes.

and is useful in NLP task. For $Affine(\cdot)$ module, we use it to integrate the characteristic information of pedestrians in a certain area. As we all know, a whole piece of map information is mostly useless for a single pedestrian, and pedestrians only focus on information in a certain area that is in line with their own direction. For example, pedestrians do not care about pedestrians far away on another sidewalk, so we use affine transformation to extract a fixed-size feature map V_p' which is consistent with the pedestrian's heading angle from the entire feature map V_p to represent the interaction area of the pedestrian, and use the attention module $ATT(\cdot)$ *to* extract the features of interest in the interaction area, then use it as useful information to decode.

We use the features $C_p(i)$ as the final encode features which include spatio-temporal information.

F. Multiple-PredictionNet

In our work, we factorize the notion of uncertainty into independent quantities, which respectively are intent uncertainty and control uncertainty. Taking the after-attention pedestrian features $C_p(i)$ as input, the Multiple-PredictionNet outputs the final behavior forecasting, which include multiple trajectories and their probabilities are corresponding with intent uncertainty and control uncertainty. In other words, for each pedestrian, model will predict K possible future trajectories and their confidence scores.

As shown in Fig. 2, the Multiple-PredictionNet has two branches, a regression branch to predict K trajectories which produces $K \times T \times 5$ parameters describing bivariate Gaussian distribution in each time step (parameterized by μ_x, μ_y, σ_x, σ_y, ρ) and a classification branch to predict confidence scores (or probability) of each trajectory which produces $K \times T \times 1$ parameters describing K softmax logits to represent $\pi(s_t|X_t)$.

$$O_{i,reg} = \{(s_{i,1}^k, s_{i,2}^k, \dots, s_{i,T}^k)\}, k \in [1, K] \tag{7}$$

$$O_{i,cls} = (p_{i,1}, p_{i,1}, \cdots p_{i,K}) \tag{8}$$

where, $s_{i,t}^k$ is bivariate Gaussian distribution parameters of regression branch output of the i^{th} pedestrian's k^{th} mode at time step t, and $p_{i,k}$ is k^{th} mode's confidence scores of i^{th} pedestrian. In order to prevent the model from collapsing during training, we will find the best K value in the next chapter.

G.Losses

In this section, we discuss the loss function that we design to train our multiple probability model. As the whole module is differentiable, so we can train our model in end-to-end way. First, given K trajectories predicted by regression predicting branch, we calculate a positive trajectory $s^{\hat{k}}$. And the $\hat{k}$ is the index of the trajectory most closely matching the ground truth trajectory s^{gt}, which measured as $\ell 2$- norm distance in state-sequence space. Then let us define a loss function $\mathcal{L}_{reg}$ of the i^{th} pedestrian's k^{th} trajectory at time step t as average displacement error (or smooth $\ell 1$ loss) between the points of positive trajectory $\hat{k}$ and other predicted trajectories. For classification, we use the max-margin loss as another loss function $\mathcal{L}_{cls}$ to Maximize the interval between positive trajectory and other predicted trajectories. Finally, we use the sum of classification and regression losses to train the model.

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha \mathcal{L}_{reg} \tag{9}$$

$$\mathcal{L}_{cls} = \frac{1}{I(K-1)} \sum_{i-1}^{I} \sum_{k \neq \hat{k}} max(0, p_{i,k} + \epsilon - p_{i,\hat{k}}) \tag{10}$$

$$\mathcal{L}_{reg} = \frac{1}{IT} \sum_{i=1}^{I} \sum_{t=1}^{T} \ell_{1_{smooth}}(s_{i,t}^{\hat{k}} - s_{i,t}^{gt}) \tag{11}$$

$$\ell_{1_{smooth}}(x) = \begin{cases} 0.5x & if\ \|x\| < 1 \\ \|x\| - 0.5 & otherwise, \end{cases} \tag{12}$$

Where $\|x\|$ denotes the $\ell 1$ norm of x, and $\alpha \in (0,1]$, ϵ is margin.

The model architecture in the article is based on Pytorch. The FCN layer in the MLP layer of the network structure is then connected to the L1 regularization layer and the activation function ReLU layer. The model uses the (2*64) embedding layer for trajectory encoding. The number of hidden layer units in the LSTM network is set to 16, and the Adam optimizer [29] is used to generate the trajectory. The minimum batch size is 64, the learning rate of the model is 0.0005, and the number of training times is 10,000.

IV. Experimental evaluation

In this section, we first introduce the experimental setting, including dataset and metric. Then we present empirical results on number of prediction tasks. Next, we compare our model with the state-of-the-art, and show great improvements in all metrics, and show the advantage of our multiple probabilistic model. Finally, we show qualitative results and discuss the future directions.

Experimental setting

A. Dataset

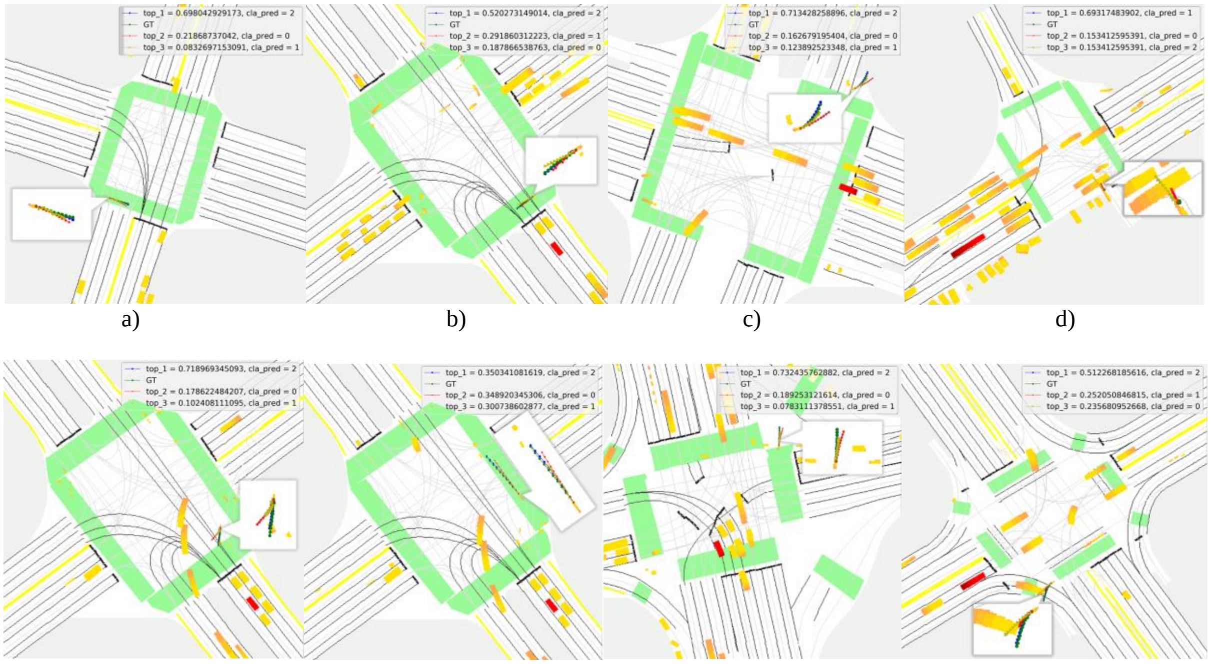
In order to verify the effectiveness of our proposed model, we report results on our in-house behavior prediction dataset. In-house dataset is a real-world driving scenes from Beijing, China. It contains HD maps data and obstacle data, these data are captured by Baidu Apollo self-driving car whose perception is equipped with lidar, cameras and radars and it can provide sufficiently accurate positions and tracks for all nearby agents, including pedestrians, vehicles and cyclists. In our dataset, the sensing vehicle is treated as an additional obstacle vehicle in the traffic scene which has no difference with other vehicles.

The total number of all agent trajectories are 800M, includeing vehicle trajectories 655M, cyclist trajectories 80M and pedestrian trajectories 65M. In our task, we need the pedestrian trajectories to verify the performance of the proposed system. The sequences have been split into training set, validation set, and test set, each of which has 45M, 10M and 10M. Each trajectory has a length of 11 seconds, where (0-3) second is the history used as observed and (4-9) second is the ground truth used for prediction. The pedestrians' trajectories are captured by the real world, including standing still, speeding through, turning, moving forward at a constant speed, walking in groups and so on. For the HD map features, our in-house dataset includes lane line, stop line, traffic light information and sidewalk information.

TABLE I

COMPARATIVE ANALYSIS OF VARIOUS BENCHMARK MODELS IN OUR IN-HOUSE DATASET METRICS

Model	time	minADE	minFDE	Recall
Linear	5s	1.43	2.98	0.15
LSTM [8]	5s	1.01	1.98	0.43
S-LSTM [10]	5s	0.89	1.72	0.51
S-GAN-20P [12]	5s	0.61	1.21	0.75
Sophie [13]	5s	0.70	1.43	0.59
Social-BiGAT [14]	5s	0.69	1.30	0.71
Social-STGCNN [16]	5s	0.71	1.35	0.82
Our Model (K=1)	1s	0.22	0.24	0.99
	3s	0.45	0.78	0.88
	5s	0.75	1.48	0.68
Our Model (K=3)	1s	**0.17**	**0.17**	**0.99**
	3s	**0.32**	**0.49**	**0.96**
	5s	**0.52**	**0.94**	**0.85**


top_1 = 0.698042929173, cla_pred = 2
GT
top_2 = 0.21868737042, cla_pred = 0
top_3 = 0.0832697153091, cla_pred = 1
top_1 = 0.520273149014, cla_pred = 2
GT
top_2 = 0.291860312223, cla_pred = 1
top_3 = 0.187866538763, cla_pred = 0
top_1 = 0.713428258896, cla_pred = 2
GT
top_2 = 0.162679195404, cla_pred = 0
top_3 = 0.123892523348, cla_pred = 1
top_1 = 0.69317483902, cla_pred = 1
GT
top_2 = 0.153412595391, cla_pred = 0
top_3 = 0.153412595391, cla_pred = 2
top_1 = 0.718969345093, cla_pred = 2
GT
top_2 = 0.178622484207, cla_pred = 0
top_3 = 0.102408111095, cla_pred = 1
top_1 = 0.350341081619, cla_pred = 2
GT
top_2 = 0.348920345306, cla_pred = 0
top_3 = 0.300738602877, cla_pred = 1
top_1 = 0.732435762882, cla_pred = 2
GT
top_2 = 0.189253121614, cla_pred = 0
top_3 = 0.0783111378551, cla_pred = 1
top_1 = 0.512268185616, cla_pred = 2
GT
top_2 = 0.252050846815, cla_pred = 1
top_3 = 0.235680952668, cla_pred = 0

a) b) c) d)

e) f) g) h)

Fig. 4. Visualization of the prediction effect of our model in several typical junction scenes of our dataset. it contains a series of typical trajectory prediction scenes, including scenes of normal pedestrian walking, accelerated walking, breaking traffic rules, and human-vehicle interaction.

B. Metrics and benchmarks

The results of the algorithm training on the dataset are compared and analyzed using various benchmarks. The experiment runs on the Ubuntu 18.04 LTS operating system, the GPU is NVIDIA Titan X, and the deep learning framework of PyTorch 1.1.0, CUDA 10.1 and Cudnn v7.5.0 are used.

Similar to the previous research methods [11], the following indicators are used to predict and evaluate the trajectory of each pedestrian, 1) Average Displacement Error (ADE) is defined as the $\ell 2$ (Euclidean) distance between the actual trajectory and the predicted trajectories sequence in each time step. 2) Final Displacement Error (FDE) is defined as the $\ell 2$ (Euclidean) distance between the actual trajectory and the predicted trajectories sequence at the final position. In this paper, because the pedestrian trajectory prediction is multi-modal by nature, we use the minimum ADE (minADE) and minimum FDE (minFDE) of the top K predictions as the metrics. And when K is equal to one, the minADE and minFDE are equal to before metrics. In addition, we use Recall to measure the stability of the predicted trajectories which is defined as the difference between the predicted point and the real point is less than 1m at 3s.

$$minADE = min\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\|s_{i,t}^{k} - s_{i,t}^{gt}\|_2\right)_{k\in[1,K]} \quad (13)$$

$$minFDE = min\left(\frac{1}{n}\sum_{i=1}^{n}\|s_{i,T}^{k} - s_{i,T}^{gt}\|_2\right)_{k\in[1,K]} \quad (14)$$

RESULT

A. Comparison with the most advanced model

This section presents empirical results on a series of prediction tasks. This paper uses the most representative Linear, LSTM, Social LSTM, Social GAN, Sophie, Social-BiGAT and Social-STGCNN as the benchmarks for comparison:

Linear, Linear regression uses the minimized MSE to optimize the parameters of the regression.

LSTM, A data-driven trajectory prediction method.

Social-LSTM, associates each pedestrian with a LSTM unit, uses the social pooling mechanism to collect and predict the hidden state of neighboring pedestrians.

Social-GAN, based on the GAN method for trajectory prediction.

Sophie, a trajectory prediction generative model based on physical and social attention.

Social-BiGAT, a trajectory prediction generation model based on graph attention network.

Social-STGCNN, a trajectory prediction model using graph network modeling and time convolution prediction.

We compare our model with the above advanced models. In order to maintain the consistency of the benchmark when evaluating the model, for the GAN model that generates a series of trajectories, the method proposed by Gupta *et al.* in [11] is used to make the model to generate *N* samples, and the generated trajectory closest to the real sample is used for evaluation, the number of data samples collected in the experiment is *N*=20. Since some models can only predict a single trajectory output, just like Linear, social LSTM and so on, in this case the ADE of the model predicted trajectory is equal to minADE, and similarly, FDE is equal to minFDE. In Table I, we can see that our model is significantly better than other models in various metrics. Comparing and analyzing each method, the prediction trajectory metrics of the model in the prediction step length increase with the increase of the prediction distance, and the accuracy and recall index of all networks have significantly decreased. It is suit for the truth that the longer trajectory prediction, the greater degree of difficulty. According to the facts, in the entire evaluation experiment, the prediction errors of each model within the prediction step have a similar trend.

As expected, because the Linear model cannot effectively model the complex social interaction between different pedestrians and understand the scene in the traffic junctions, it performs the worst in the all trajectory prediction tasks. With the use of LSTM in sequence prediction filed, the accuracy of model prediction has gradually improved, it also proved that the effectiveness of sequence data for predicting problems. Compared with S-LSTM, whose interaction pooling model is based on grid division, the S-GAN model tends to generate multiple random trajectories, which has higher accuracy and social acceptability than the previous generation of a deterministic trajectory. But GAN network is very hard to train, especially when the trajectory generator and the trajectory discriminator are unbalanced, it is easy to cause the gradient vanishing problem or mode collapse (the sampled synthetic data has no diversity), especially when predicting pedestrian trajectory, avoiding mode collapse is essential for autonomous driving decision-making and driving safety. Social-STGCNN [16] adopted a completely different method, uses graph networks to model crowd interactions, and uses temporal convolution network instead of recursive loop architecture. But none of the above models consider the spatial topology in the scene, this largely limits the performance of the model as well.

Similar to our work, Social-BiGAT [14] also uses semantic scene information and sequence information as input. But all metrics of our model are better than this model. This is because we use a composite raster map instead of directly inputting the picture from sensors into the model, our model can effectively learn the map topology and the complex interaction among traffic agents. In their works, the pedestrians' state is encoded independently so the connection between the global map topology and single pedestrian cannot be captured exactly. We present an attention module which can concatenate the scene feature map and the pedestrian sequence feature by multi-head attention mechanism. So, our model can excellently model semantic features and employ multi-head attention mechanism to capture the feature is more accurate and useful.

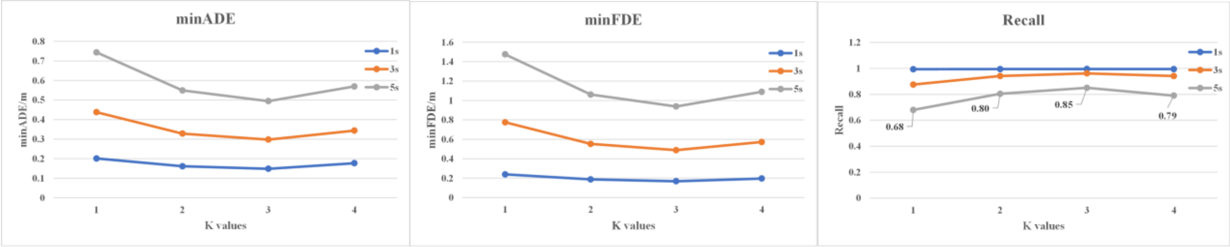
Fig. 4 shows the probabilistic visualization maps of multiple predicted trajectories in the traffic junction scene of our dataset. In this part, we will show our model's ability to accurately predict the future trajectories of pedestrians when dealing with different scenarios, traffic conditions, and emergency situations. As shown in Fig. 4-(a, b, c), the model predicts the reliability of the trajectory in different junctions. Our model predicts 3 trajectories of each pedestrian which include top1, top2, and top3, the trajectory of top1 is the predicted trajectory with the highest probability value, which is most consistent with Ground Truth (GT). In Fig. 4-(d), a pedestrian is in a T-shape junction with dense traffic, the top1 trajectory with 69% probability predicted by the model is consistent with GT, and the state of pedestrian is waiting at this moment. This situation is especially common in autonomous driving tasks. For autonomous vehicles, the ability to accurately predict the location and intention of pedestrians when driving on the road is very important for road traffic safety and ease of traffic pressure, especially in dense traffic scenes. Fig. 4-(e, f) shows that our model can predict special situation accurately that a pedestrian is violating traffic rules which require pedestrians to walk on sidewalk, and quickly cross the sidewalk in Fig. 4-(f). The next picture shows that the model is integrating the information of the semantic map. It is noted that the sidewalk on the right side, not only the trajectory of top1 can be well matched with GT,

but also the trajectory of top2 and top3 are consistent with the intention of pedestrians in the future which is going to the next sidewalk on the right side. Finally, in the last graph, we can see our model can deal with the interaction with the vehicle well.

TABLE II

COMPARATIVE ANALYSIS OF MODEL METRICS WITH DIFFERENT K VALUES.

K	Forecast duration	minADE	minFDE	Recall
1	1s	0.22	0.24	0.99
	3s	0.45	0.78	0.88
	5s	0.75	1.48	0.68
2	1s	0.18	0.19	0.99
	3s	0.35	0.55	0.94
	5s	0.56	1.06	0.80
3	1s	**0.17**	**0.17**	**0.99**
	3s	**0.32**	**0.49**	**0.96**
	5s	**0.52**	**0.94**	**0.85**
4	1s	0.18	0.20	0.99
	3s	0.34	0.57	0.94
	5s	0.57	1.09	0.79



a) b) c)

Fig. 5. Visualization diagram of comparative analysis of model metrics with different K values.

B. Quantitative analysis: the choice of K value

In this section, we analyze the influence of the choice of different prediction K values on the model prediction accuracy.

In Table II, the metrics are minADE, minFDE, and Recall for K = 1, K = 2, K = 3, and K = 4 in 1 second, 3 second and 5 second. As shown in Table II and Fig. 5, when the model takes different values of K, the model has different accuracy, and a value of 3 corresponds to the highest model metrics. Observation can be drawn from the result. K represents the number of behaviors that pedestrians can take in future. Different K corresponds to different behavior patterns. Intuitively, when K is equal to 3, our model can model trajectory problem very well. In other experiments of the article, we selected three branches as prediction result of multiple probability prediction model.

V. CONCLUSION

In this paper, we propose a novel pedestrian trajectory prediction model to learn the complex interactions between pedestrians and surroundings in traffic junctions. In order to learn the topological relationship of different features in the traffic scene, we first design a composite raster map, which has all the features needed to make a reasonable prediction. Then we design AgentNet and MapNet to process the temporal and spatial information in semantics and sequence features. Next, we use Interaction-AttNet to aggregate the features through the attention module and pass them to PredictionNet to predict multiple probability trajectories. We conduct experiments on the large scale in-house dataset and the results show that our method is significantly outperforms the state-of-the-art. A natural next step is to extend the pedestrian multiple probability attention model in the junction scene to predict the trajectories and intentions of all moving agents in this scene.

REFERENCES

[1] A. Rudenko, “Human motion trajectory prediction: a survey,” *Int. J. Robot Res.*, vol. 39, no. 8, pp. 895–935, Jul. 2020.

[2] M. Moussaid, “The walking behaviour of pedestrian social groups and its impact on crowd dynamics,” *Plos one*, vol. 5, no. 3, Apr. 2010.

[3] Y.-N. Chai *et al.*, “MultiPath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction,” 2019, *arXiv:1910.05449*.

[4] N. Schneider and D. Gavrila, “Pedestrian path prediction with recursive Bayesian filters: a comparative study,” in *Proc. German Conference on Pattern Recognition*, Saarbrucken, Germany, 2013, pp. 174–183.

[5] K. Kim, “Gaussian process regression flow for analysis of motion trajectories,” in *Proc. IEEE Int. Conf. Comput. Vision.*, Barcelona, Spain, 2011, pp. 1164-1171

[6] K. M. Kitani and B. D. Ziebart, “Activity Forecasting,” in *Proc. Eur. Conf. Comput. Vision*, Florence, Italy, 2012, pp. 201–214.

[7] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, “Who are you with and where are you going?” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, Jun. 2011, pp. 1345–1352.

[8] S. Hochreiter and J. Schmidhuber, ‘‘Long short-term memory,’’ *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[9] A. Vemula, K. Muelling, and J. Oh, ‘‘Social attention: Modeling attention in human crowds,’’ in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2018, pp. 1–7.

[10] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, ‘‘Social LSTM: Human trajectory prediction in crowded spaces,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 961–971.

[11] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, ‘‘Social GAN: Socially acceptable trajectories with generative adversarial networks,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2255–2264.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, ‘‘Generative adversarial nets,’’ in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[13] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, ‘‘SoPhie: An attentive GAN for predicting paths compliant to social and physical constraints,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 1349–1358.

[14] V. Kosaraju, A. Sadeghian, and R. Martin-Martin, “Social-BiGAT: multimodal trajectory forecasting using bicycle- GAN and graph attention networks,” 2019, *arXiv:1907.03395*.

[15] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Li, and Y. Bengio, “Graph attention networks,” 2018, *arXiv:1710.10 903*.

[16] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, “Social-STGCNN: asocial spatio-temporal graph convolutional neural network for human trajectory prediction,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun, 2020, pp. 14412–14420.

[17] Z.-X. He, C.-Y. Chow, and J.-C. Zhang, “STCNN: A spatio- temporal convolutional neural network for long-term traffic prediction,” in *Proc. IEEE 12th Int. Conf. Mobile Data Manage,* Hong Kong, China, 2019, pp. 226–233.

[18] G.-H. Li, M. Mueller, A. Thabet, and B. Ghanem, “DeepGCNs: Can GCNs go as deep as CNNs?” in *Proc. IEEE Int. Conf. Comput. Vis.*, Seoul, South Korea, 2019, pp. 9266–9275.

[19] S. Casas, C. Gulino, and R. Urtasun, “The importance of prior knowledge in precise multimodal prediction” 2020, *arXiv:2006.02636*.

[20] H.-G. Cui, *et al.*, “Multimodal Trajectory Predictions for Autonomous Driving using Deep Convolutional Networks,” in *Proc. IEEE Int. Conf. Robot. Autom*, Montreal, QC, Canada, 2019, pp. 2090–2096.

[21] H. Cheng, W.-T. Liao, *et al.*, “MCENET: Multi-context encoder network for homogeneous agent trajectory prediction in mixed traffic,” 2020, *arXiv:2002.05966*.

[22] B. Yang, G. Yan, *et al.*, “TPPO: A Novel Trajectory Predictor with Pseudo Oracle,” 2020, *arXiv:2002.01852*.

[23] J. Liang, L. Jiang, J. C. Niebles, A. G. Hauptmann, and L. Fei-Fei, ‘‘Peeking into the future: Predicting future person activities and locations in videos,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5725–5734.

[24] V. Mnih, N. Heess, and A. Graves, “Recurrent models of visual attention” in *Proc. Adv. Neural Inf. Process. Syst*., Montreal, Canada, 2014, pp. 2204–2212.

[25] K. Xu, J. Ba, *et al*., “Show, attend and tell: Neural image caption generation with visual attention,” 2015, *arXiv:1502.03044*.

[26] D. P. Kingma, M. Welling, “Auto-Encoding variational bayes,” 2013, *arXiv:1312.6114*.

[27] A. Vaswani, N. Shazeer, *et al*., “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst*., Long Beach, CA, USA, 2017, pp. 5998–6008.

[28] Y. Hoshen, “VAIN: Attentional multi-agent predictive modeling,” 2017, *arXiv:1706.06122*.

[29] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, *arXiv:1412.6980*.

[30] N. Rhinehart, R. McAllister, K. Kitani, and S. Levine, “PRECOG: Prediction conditioned on goals in visual multi-agent settings,” in *Proc. IEEE Int. Conf. Comput. Vision.*, Seoul, Korea (South), Oct. 2019, pp. 2821–2830.

Linhui Li received the Ph.D. degree in Vehicle Operation Engineering from Jilin University, Jilin, China, in 2008. He is an associate professor at Dalian University of Technology. His main research interest lies in intelligent vehicle trajectory prediction, vision based environmental perception of intelligent vehicle, and navigation control. He was a visiting scholar at The Ohio State University from 2017 to 2018, and is the author of over 40 publications.

Bin Zhou received the B.S. degree in 2018 from Shandong University of Science and Technology, China. He is working toward a M.S. degree at Dalian University of Technology, Dalian, China. His research interests include decision-making and trajectory prediction of autonomous vehicles.

Jing Lian received the Ph.D. degree in Communication and Information System from Jilin University, Jilin, China, in 2008. She is an associate professor and the deputy director of Automotive Electronic Institute at Dalian University of Technology, the judicial expert in the vehicle performance. Her main research interest lies in new energy vehicle intelligence, trajectory prediction, and motion planning of autonomous vehicles. She is the leader of more than 20 research projects, and is the author of over 60 publications.

Xuecheng Wang received the B.S. degree in 2020 from Henan Polytechnic University, China. He is working toward a M.S. degree at Dalian University of Technology, Dalian, China. His research interests include trajectory prediction of autonomous vehicle.

Yafu Zhou received the B.S. degree and the M.S. degree from Tianjin University, China in 1986 and 1989, respectively. He is a professor at the School of Automotive Engineering of Dalian University of Technology, a project review expert of Ministry of Science and Technology of China, a science and technology awards evaluation expert of Ministry of Education of China. His research interests include intelligent vehicles and power control of new energy vehicles. He has published more than 60 academic papers, and more than 20 authorized national invention patents.