Intention-Aware Vehicle Trajectory Prediction Based on Spatial-Temporal Dynamic Attention Network for Internet of Vehicles

Xiaobo Chen, *Member, IEEE*, Huanjia Zhang, Feng Zhao, Yu Hu, Chenkai Tan, and Jian Yang, *Member, IEEE*

***Abstract*— Vehicle trajectory prediction is a keystone for the application of the internet of vehicles (IoV). With the help of deep learning and big data, it is possible to understand the between-vehicle interaction pattern hidden in the complex traffic environment. In this paper, we propose a novel spatial-temporal dynamic attention network for vehicle trajectory prediction, which can comprehensively capture temporal and social patterns in a hierarchical manner. The social relation between vehicles is captured at each timestamp and thus retains the dynamic variation of interaction. The temporal correlation in terms of individual motion state as well as social interaction is captured by different sequential models. Furthermore, a driving intention-specific feature fusion mechanism is proposed such that the extracted temporal and social features can be integrated adaptively for the maneuver-based multi-modal trajectory prediction. Experimental results on two real-world datasets show that compared with the state-of-the-art algorithms, our proposal achieves comparable prediction performance for short-term prediction, however, works much better for long-term prediction. Additionally, various ablation analysis is provided to evaluate the effectiveness of our proposed network components. The code will be available at https://xbchen82.github.io/resource/.**

***Index Terms*— Trajectory prediction, multi-head self-attention, multi-modal prediction.**

I. Introduction

DUE to the rapid development of communication networks such as C-V2X and 5G, the high transmission bandwidth and low latency have paved the way for the Internet of Vehicles (IoV) [1] where vehicles can share their local information including but not limited to the ego-motion state, road environmental data captured by different types of sensors, such as camera, LiDAR, to name a few. Especially, the vehicle trajectory data can be gathered by Mobile Edge Computing (MEC) server where real-time trajectory prediction can be carried out. The accurate prediction of vehicle trajectory [2] in certain traffic scenes has been an indispensable task with promising applications, e.g., autonomous driving. Motion trajectory data [3] often contain abundant spatial and temporal information that should be analyzed for trajectory prediction. In the rest of this paper, the vehicle whose trajectory is to be predicted is referred to as the target vehicle whereas the other vehicles within a certain distance of the target vehicle are called surrounding vehicles.

There are many challenges for accurately modelling and predicting the future trajectory of the target vehicle due to its complex nature. In the dynamic and dense driving environment, the motion of a vehicle is often influenced by both its own history path and the interactions with neighboring vehicles. For instance, if the target vehicle wants to make a lane change, it needs to consider the distance to the other vehicles in the target lane so as to avoid collisions. As a result, it becomes rather important to analyze the influence of the surrounding vehicles on the target vehicle we want to predict. Unfortunately, the potential interaction pattern between vehicles is unknown and has to be inferred from the raw trajectory data.

In addition, there might be multiple plausible trajectories the target vehicle can take even given the same driving situation because it is difficult to have a comprehensive and correct understanding of other unknown factors affecting driving behavior, such as personalized driver characteristics [2], physical and psychological factors [4], etc. For instance, as shown in Fig. 1, under the same situation, the target vehicle has the option to keep its lane or change to another lane. Even when doing the same maneuver, the execution can have differences in terms of operation pattern. Therefore, how to model such multi-modality of driving intention is essential for the generation of multiple possible trajectories.

To address the above challenges, significant progress has been made towards vehicle trajectory prediction during the last few years. Earlier works generally adopted traditional machine learning approaches, including Bayesian learning,

Manuscript received 2 October 2021; revised 15 February 2022 and 31 March 2022; accepted 6 April 2022. Date of publication 3 May 2022; date of current version 11 October 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61773184 and Grant 62176140, in part by the National Key Research and Development Program of China under Grant 2018YFB2100500, in part by the Taishan Scholars Program of Shandong Province under Grant tsqn202103118, and in part by the Six Talent Peaks Project of Jiangsu Province under Grant 2017-JXQC-007. The Associate Editor for this article was W. Lin. *(Corresponding author: Xiaobo Chen.)*

Xiaobo Chen, Huanjia Zhang, and Feng Zhao are with the School of Computer Science and Technology, Shandong Technology and Business University, Yantai, Shandong 264005, China (e-mail: xbchen82@gmail.com; 1036882112@qq.com; zhaofeng1016@126.com).

Yu Hu and Chenkai Tan are with the Automotive Engineering Research Institute, Jiangsu University, Zhenjiang, Jiangsu 212013, China (e-mail: jiongger@gmail.com; tanchenkai@163.com).

Jian Yang is with the College of Computer Science, Nankai University, Tianjin 300350, China (e-mail: csjyang@nankai.edu.cn).

Digital Object Identifier 10.1109/TITS.2022.3170551

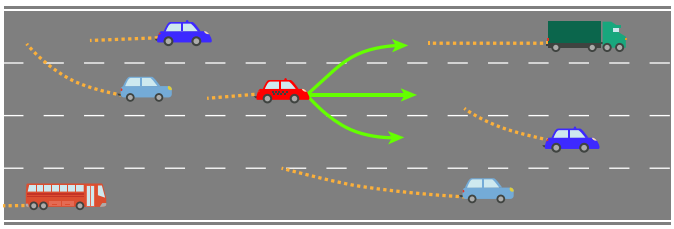


Fig. 1. Illustration of multi-modality trajectory prediction.

hidden Markov models (HMMs), support vector machines (SVMs), Gaussian Process (GP) [5], etc., for trajectory prediction. Those methods generally need to extract hand-crafted features [6] from the raw trajectory data before applying specific recognition model. However, it is unrealistic to design a universal feature representation manually which is suitable for a variety of traffic environments because of different scene contexts. Consequently, traditional methods are difficult to obtain satisfactory accuracy especially for the long-term prediction (3-5 seconds) which is proved to be more challenging than short-term prediction (1-3 seconds) [8].

Recently, benefiting from the development of deep learning techniques and the availability of big real-world trajectory data, various data-driven trajectory prediction methods based on deep neural networks have been developed, which significantly push forward the research to a new level. Trajectory data can be viewed as interactive multivariate timeseries, and therefore capturing the temporal and interaction dependency is one of the key steps towards accurate prediction. In order to capture the temporal correlation of trajectory between different timestamps, recurrent neural networks (RNNs) and especially a series of improved variants, such as long short-term memory network (LSTM) [9], gated recurrent unit (GRU) [10], BiRNN [11], are widely adopted as building blocks because of their capability to model sequential data. In order to characterize the social interaction, convolutional neural network (CNN) has been applied to the traffic scene which is divided into 2D occupancy grids with fixed size [12]. Besides CNN, graph neural network (GNN) [13] particularly suitable for modelling the potential correlation between multiple spatially distributed nodes is leveraged to express the relationship between vehicles in the same scene. More recently, due to the great success achieved by the transformer-based models [14], more studies begin to explore the feasibility of applying transformer to address the challenges in trajectory prediction [15]–[17]. Different from CNN and LSTM, the transformer network is constructed purely based on multi-head attention mechanism and stacking layers which enables it to learn dynamic and hierarchical features including the long-term correlations and multidimensional dynamic dependency in sequential data. These characteristics make the transformer architecture a natural selection for sequence-to-sequence (seq2seq) problem [14].

In order to achieve accurate vehicle trajectory prediction, we should make full use of the social and temporal information contained in the historical trajectory. Intuitively speaking, the short-term prediction depends closely on the motion and interaction context at the latest few timestamps, however, for long-term prediction, the spatial and temporal information across the whole historical trajectory needs to be fully exploited. Generally speaking, the more detailed historical information we obtain, the more spatial and temporal features we can extract from traffic context, thus promoting the understanding of long-term motion and social regularity of vehicles. To this end, learning feature representation containing abundant information from raw trajectory data is rather crucial for accurate prediction. Accompanied by this problem, we need an effective mechanism to filter and fuse the social and temporal features extracted at different timestamps which may contain much redundant information with respect to a specific task. Inspired by the recently emerging multiple knowledge representation (MKR) theory which highlights the complementary capacity of multiple streams of information (or multiple representations) [18], [19], we are devoted to a novel network architecture that can characterize and integrate social and temporal representations to facilitate the mutual benefits of each other hierarchically.

Based on the above discussion, this paper proposes a spatial-temporal dynamic attention network for intention-aware vehicle trajectory prediction. This model is comprised of several specialized modules organized in a hierarchical manner to capture different levels of social and temporal features from low-level motion pattern to high-level semantic representation. We propose a multi-head attention-based temporal and social dependency modelling approach which can extract abundant features from raw trajectory data at different timestamps. The high-level dynamic dependency of social interaction also can be captured. Moreover, a feature fusion method devoted to different driving maneuvers is proposed such that different combinations of features can be generated aiming at intention-aware trajectory prediction. Each module of our proposed model is validated by a series of ablation studies. The major contributions of this work are summarized as follows.

- The social interaction at different timestamps is captured through multi-head attention mechanism which effectively measures the influence of surrounding vehicles on the target vehicle. By doing so, the model has the capability to retain the social interaction features at all historical timestamps.
- The temporal dependency between the social interaction across consecutive timestamps is modelled by using multi-head self-attention mechanism. Different from low-level motion correlation extracted from raw trajectory features, such high-level temporal correlation in terms of social interaction also plays an important role and can be beneficial when understanding the motion of interacting vehicles.
- A novel intention-specific feature fusion method is proposed to aggregate the temporal and interaction features from different timestamps for joint intention recognition and trajectory prediction. This method achieves multi-modal prediction by directly bridging inherent intention and possible trajectories through intention-specific feature combinations.
- Experiments on real-world vehicle trajectory datasets are carried out to evaluate the proposed method

comprehensively. The results and ablation study are provided to verify the effectiveness of our method. The experiments show that compared with the state-of-the-art models, our method can achieve comparable performance for short-term prediction (1-3 seconds), however, works much better in terms of long-term prediction (3-5 seconds).

The rest of this paper is organized as follows. Section II presents a literature review of related works towards vehicle trajectory prediction. Section III shows the formulation of the vehicle trajectory prediction problem and presents the architecture of our proposed spatial-temporal dynamic attention network. Section IV reports the experimental results and ablation analysis and finally, Section V concludes this paper.

II. Related Works

For vehicle trajectory prediction, two crucial components can be identified which are multi-agent interaction learning and multi-modality distribution learning. The former models the social and temporal influence among agents while the latter predicts multiple possible trajectories. In this section, we briefly discuss some existing studies on these two problems.

A. Multi-Agent Interaction Learning

The movement of a vehicle is not only influenced by its past trajectory but also related to other neighboring vehicles especially in the dense traffic environment. Therefore, capturing the relevance between vehicles plays a vital role in trajectory prediction. Until now, there have been a lot of works for modeling social interaction for either vehicle or pedestrian prediction. Some works further incorporate environment features such as obstacles [20]–[22] to reflect their impact for more accurate prediction. A classic study proposed early is the social force model (SFM) [23] which reflects the interaction between the pedestrians by attractive and repulsive forces. However, this model is manually designed and thus lacks enough flexibility to reflect the complex characteristics of real traffic environment. Recently, deep learning has been extensively applied for trajectory prediction due to the prominent feature learning capability of multi-layer neural networks. Firstly, raw trajectory data comprising both temporal and spatial information need to be represented in an appropriate form before feeding it to neural networks. Typical representation includes sequential points [16], [17], occupy girds [12], [25], rasterized image [26], etc. Based on these representations, a variety of network components can be used to extract temporal and social context. For the temporal correlation, either recurrent or convolutional network can be used with respective advantages and disadvantages. For instance, recurrent networks, e.g., RNN, LSTM, GRU, are rather suitable for sequential data due to their memory mechanism. However, they may suffer from slow training and inference speed as well as vanishing gradients problem. A holistic LSTM [27] was recently proposed which introduces speed, intention, and correlation memory cells into canonical LSTM to improve the transferability in modeling future variations of pedestrian trajectory. In addition, convolutional networks, such as temporal convolutional network (TCN) [28], [29], also can be used to model timeseries data with much less time complexity however has to contain more layers when capturing the long-range correlation. For the inter-vehicle interaction, social pooling [16], [30], convolutional social pooling [12], [31], tensor fusion [28], [32], and graph neural network [33]–[35] are some popular candidates. Recently, due to the rapid development of transformer model, various methods based on attention [17], [24], [36] have been proposed. Modelling social interaction between vehicles by attention has an obvious advantage that dividing traffic scene into occupancy grids [12], [25] is not required thus eliminating manual intervention. In this work, we attempt to exploit the interaction information contained in all historical timestamps and further capture the potential dynamic interaction dependency across different timestamps. The social interaction at each timestamp as well as its dynamic evolution is modelled by attention.

B. Multi-Modal Distribution Learning

Due to the diversity and invisibility of hidden maneuvers, the vehicle can have multiple feasible options to conduct even given the same traffic context information. Therefore, it is beneficial to predict multiple possible trajectories, otherwise, the average of several modes may be produced thus causing mode collapse. Such multi-modality is a unique characteristic of pedestrian or vehicle movement [37]–[39] and often characterized by introducing extra latent variables. The existing models to address this issue can be categorized into two classes based on whether the latent variable has concrete semantics or interpretability. In the first class, the latent variable has clear semantics. For instance, some studies [12], [17], [31] treat driving maneuver as a latent variable which is classified firstly and then used to build a Gaussian mixture model for multi-modal trajectory prediction. Similarly, trajectories can be divided according to the pre-clustered anchors [40] rather than the predefined maneuvers. Other works [41] feed different lane features in the decoder to generate the corresponding prediction. However, most studies directly apply a simple concatenation operation to combine the features of maneuver or lane and encoded context which are not powerful enough for generating diverse predictions. In the second class, the latent variable lacks concrete semantics. These models are generally motivated by the recent progress of generative deep learning models such as variational autoencoder (VAE) [42] and generative adversarial network (GAN) [32], [43], [44] with the capability of generating complex distribution. These methods are generally based on the injection of the noise sampled from latent distribution. The noise along with the encoded contextual features is decoded to produce stochastic predictions. The model is learned by optimizing the low-bound of the likelihood function or an adversarial loss. Besides the lack of specific semantic interpretation, it is difficult to determine the number of samples that can cover all possible trajectories in practice. Moreover, it has to address the issue assigning proper probabilities to the randomly generated trajectories. Our work belongs to the first class because the

latent variables in the proposed model exactly correspond to the possible driving intentions with clear semantics. More importantly, we propose an intention-specific feature fusion module where different driving intentions will correspond to different feature combination matrices. By doing so, this module implements multi-modal prediction which is not only flexible in feature aggregation but also leads to improved performance.

III. Methodology

A. Problem Formulation

In this paper, the vehicle trajectory prediction problem is formulated as the prediction of the probability distribution of the target vehicle future trajectory position based on the observed historical motion information of the target vehicle and its surrounding vehicles. Formally, let $X_t = \left\{x_t^0, x_t^1, x_t^2, \cdots, x_t^N\right\}$ be the states of total $N+1$ vehicles at timestamp t, the state x_t^i associated with vehicles i could include its x and y coordinates, velocity, acceleration, vehicle type, etc. In this paper, superscript refers to vehicle while subscript refers to timestamp. Without loss of generality, let x_t^0 and x_t^i $(i \geq 1)$ be the state of the vehicle being predicted and the surrounding vehicle i, respectively. Let $Y = \left\{y_{T+1}^0, y_{T+2}^0, \cdots, y_{T+F}^0\right\}$ denote the predicted trajectory of the target vehicle from timestamp $T+1$ to $T+F$, F stands for the prediction horizon, y_{T+f}^0 $(1 \leq f \leq F)$ often consists of the x and y coordinates of the target vehicle at future timestamp $T+f$. Notice that the coordinates of all vehicles are expressed in a reference frame where the origin is the position of the target vehicle at timestamp T.The input to the model consists of $X = \{X_1, X_2, \cdots, X_T\}$ during the past T timestamps and the output of the model is a probability distribution $P(Y|X)$ over Y. In this work, the distribution of y_{T+f}^0 is parameterized as a bivariate Gaussian distribution with mean $\left(\mu_{T+f,x}, \mu_{T+f,y}\right)$, variance $\left(\sigma_{T+f,x}^2, \sigma_{T+f,y}^2\right)$, and correlation coefficient ρ_{T+f}.

B. Model Architecture

To achieve precise trajectory prediction in dense traffic, it is essential to capture the complex temporal and social correlations between the target vehicle and other vehicles. Therefore, we propose a hierarchical spatial-temporal dynamic attention network (STDAN) model consisting of five dedicated components as shown in Fig. 2. ① **Motion Encoder (ME)** module, where the low-level motion state of all vehicles is encoded by an LSTM encoder thus extracting the temporal (sequential) relation from raw trajectory data. ② **Social Interaction (SI)** module, which aims to capture the social interaction between the target vehicle and other vehicles at different timestamps. ③ **Dynamic Interaction Dependency (DID)** module, which aims to directly capture the long-range temporal dependence in the social interaction representation sequence. ④ **Intention-specific Feature Fusion (IFF)** module where the social and temporal features from different timestamps are elaborately fused based on different driving maneuvers such that sufficient and effective information can be retained. ⑤ **Multi-modal Trajectory Prediction (MTP)** module, which implements multi-modal trajectory prediction for future timestamps based on the fused feature associated with different maneuvers.

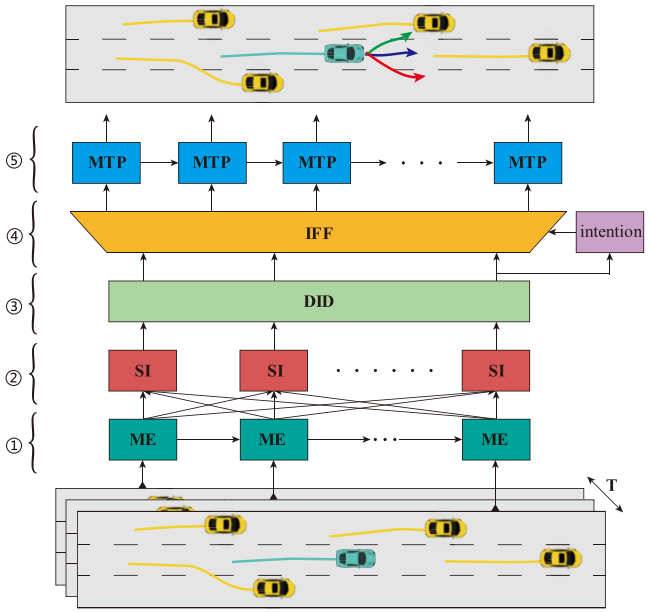


Fig. 2. Illustration of the proposed STDAN model.

1) Motion Encoder (ME) Module: This module is used to embed and encode the trajectories of all the vehicles in the scene at each observation time. Firstly, a fully connected layer is used to convert the input motion state x_t^i of vehicle i at timestamp t to an embedding representation.

$$e_t^i = \mathrm{MLP}\left(x_t^i, W_{emb}\right) \tag{1}$$

where e_t^i is the embedded feature vector, MLP denotes multi-layer perception with parameter W_{emb} and Exponential Linear Unit (ELU) activation function [45] shown below.

$$\mathrm{ELU}(x) = \begin{cases} x, & \text{if } x \geq 0 \\ \alpha\left(e^x - 1\right), & \text{if } x < 0 \end{cases} \tag{2}$$

When the input is larger than zero, ELU activation would act as an identity function while when the input is much smaller than zero, ELU activation would generate a constant. Then, LSTM cells are utilized to encoder e_t^i to take into the temporal correlation

$$h_t^i = \mathrm{LSTM}\left(e_t^i, h_{t-1}^i, W_{en-lstm}\right) \tag{3}$$

where $W_{en-lstm}$ denotes the parameter matrix of encoding LSTM cells to be learned, h_t^i denotes the hidden state vector of vehicle i at timestamp t. Notice that the LSTM operation is applied to each vehicle individually and the parameters of encoding LSTM are shared among all vehicles in order to address the issue caused by a varying number of vehicles in different scenes. Finally, for the target vehicle and vehicle i, we acquire the corresponding feature representation $\left[h_1^0, h_2^0, \cdots, h_T^0\right]$ and $\left[h_1^i, h_2^i, \cdots, h_T^i\right]$ where $h_t^0, h_t^i \in R^{d_e}$, d_e is size of the hidden state of the encoding LSTM.

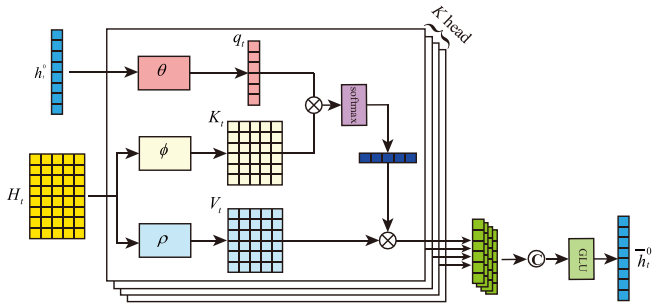


Fig. 3. Illustration of SI module at timestamp t.

2) Social Interaction (SI) Module: Obviously, an experienced driver can adjust the vehicle state including speed, heading yaw, etc., according to the behavior of its neighboring vehicles. Such social dependency could be beneficial for vehicle trajectory prediction, especially in dense traffic. In the above ME module, vehicles are regarded as independent of each other and the relation between the target vehicle and its neighboring vehicles is neglected. Therefore, it is highly necessary to capture the social relations between the target vehicle and the others in the same scene and here where the SI module comes to play. Specifically, at timestamp t, given the input feature of all neighboring vehicles denoted by $H_t = \left[h_t^1, h_t^2, \cdots, h_t^N\right]$, the feature of the target vehicle h_t^0, both of which are acquired by the above ME module, the result of SI module at time t is $\overline{h}_t^0$ according to Fig. 3.

In particular, SI module is based upon the multi-head scaled dot-product attention architecture. We first use three learnable linear transformations θ, ϕ, ρ to obtain three different vectors

$$\begin{cases} q_t = \theta\left(h_t^0, W_\theta^{so}\right) \\ K_t = \phi\left(H_t, W_\phi^{so}\right) \\ V_t = \rho\left(H_t, W_\rho^{so}\right) \end{cases} \tag{4}$$

where $q_t \in R^{d_q}$, $K_t = \left[k_t^1, k_t^2, \cdots, k_t^N\right] \in R^{d_k \times N}$, $V_t = \left[v_t^1, v_t^2, \cdots, v_t^N\right] \in R^{d_v \times N}$ are query, key, and value, respectively. W_θ^{so}, W_ϕ^{so}, and W_ρ^{so} are weight matrices that will be learned for characterizing social interaction. Then, the dot product is used to obtain attentive weight α_t^i for each pair $\left(h_t^0, h_t^i\right)$

$$\alpha_t = \text{softmax}\left(\langle q_t, K_t\rangle / \sqrt{d_k}\right) \tag{5}$$

where $\langle \cdot, \cdot \rangle$ denotes dot product operation, $\alpha_t = \left[\alpha_t^1, \alpha_t^2, \cdots, \alpha_t^N\right]$ is the attentive scoring vector with each entry α_t^i indicating how much surrounding vehicle i is relevant to the target vehicle at timestamp t. Subsequently, according to the obtained attentive weight α_t^i, the features of neighboring vehicles will be aggregated to the target vehicle by

$$head_t = \sum_{i=1}^{N} \alpha_t^i v_t^i \tag{6}$$

Similar to the vanilla multi-head attention, we use K independent social attention described above to retain different kinds of social relationships. The use of multi-head attention is originally devised by the Transformer architecture aiming to enhance the capability of modelling complex relations from multiple perspectives. Formally, let $head_{t,1}, head_{t,2}, \cdots, head_{t,K}$ be the feature representations obtained by K independent social attentions. These features are concatenated together to form $Head_t = \left[head_{t,1}, \cdots, head_{t,K}\right]$. In order to suppress the redundant features and meanwhile retain the relevant features, we apply the following Gated Linear Units (GLUs) [46] to extract complex social dependency between the target vehicle and other vehicles

$$\overline{h}_t^0 = \text{LayerNorm}\left(\text{GLU}\left(Head_t, W_{GLU}^1\right) + h_t^0\right) \tag{7}$$

where LayerNorm denotes Layer Normalization [47] and GLU takes the form

$$\begin{aligned} \text{GLU}\left(\gamma, W_{GLU}\right) = \sigma\left(W_{GLU}\gamma + b_{GLU}\right) \\ \odot \left(W_{GLU}\gamma + b_{GLU}\right) \end{aligned} \tag{8}$$

Here, σ denotes the sigmoid activation function, $\odot$ is the element-wise Hadamard product. Therefore, GLU provides the model the flexibility to select relevant interaction features from $Head_t$ adaptively. As can be seen from (7), we also add the shortcut connection [48] to enable efficient information flow across different layers. Finally, the above operation is carried out for each timestamp t, thus yielding a sequence $\overline{H}^0 = \left[\overline{h}_1^0, \overline{h}_2^0, \cdots, \overline{h}_T^0\right] \in R^{d_v \times T}$ which is the output of SI module.

3) Dynamic Interaction Dependency (DID) Module: The above social interaction feature is calculated for each timestamp t separately without considering the temporal correlation among these social representations. However, for the vehicle trajectory prediction problem, the social dependency at different timestamps could be temporally correlated. Thus, after SI module, we propose a dynamic interaction dependency (DID) module based on multi-head self-attention to capture the relationship between the social interaction representations across different timestamps. This is shown in Fig. 4.

Specifically, DID module takes the output sequence $\overline{H}^0 = \left[\overline{h}_1^0, \overline{h}_2^0, \cdots, \overline{h}_T^0\right]$ from SI module as the input. Following self-attention mechanism, the query, key, and value matrices are calculated as follows:

$$\begin{cases} \overline{Q}^0 = \theta\left(\overline{H}^0, W_\theta^{di}\right) \\ \overline{K}^0 = \phi\left(\overline{H}^0, W_\phi^{di}\right) \\ \overline{V}^0 = \rho\left(\overline{H}^0, W_\rho^{di}\right) \end{cases} \tag{9}$$

where W_θ^{di}, W_ϕ^{di}, and W_ρ^{di} are the learnable parameter matrices for the corresponding transformation. Then, the temporal self-attention is represented as

$$\beta = \text{softmax}\left(\langle \overline{Q}^0, \overline{K}^0 \rangle / \sqrt{\overline{d}_k}\right) \tag{10}$$

where $\beta \in R^{T \times T}$ is the attentive scoring matrix with entry β_{ij} measuring the temporal correlation between timestamp i

and j. Then, this scoring matrix is used to aggregate temporal information from the corresponding values

$$\overline{head} = \beta \overline{V}^0 \tag{11}$$

Similar to the above SI module, multi-head self-attention is used to jointly aggregate information from different representation subspaces thus enhancing the representation capability of the model. The obtained representations are combined and then projected through GLU to extract beneficial information and suppress the irrelevant features. Additionally, shortcut connection and layer normalization are used to facilitate the model training. The above process is expressed as

$$\tilde{h}_t^0 = \text{LayerNorm}\left(\text{GLU}\left(\overline{Head}, W_{GLU}^2\right) + \overline{H}^0\right) \tag{12}$$

where $\overline{Head} = \left[\overline{head}_1, \overline{head}_2, \cdots, \overline{head}_K\right]$ denotes the results of K-head self-attention. Finally, the output of DID module is denoted as $\tilde{H}^0 = \left[\tilde{h}_1^0, \tilde{h}_2^0, \cdots, \tilde{h}_T^0\right]$ characterizing the temporal dependency of the social interaction across different timestamps. The calculation performed by DID module is illustrated in Fig. 4

Notice that in this work, both ME module and DID module are used to capture the temporal dependency, however, they capture the feature from different perspectives and with different techniques. The former focuses on the extraction of temporal relations of raw trajectory data by an encoding LSTM while the latter captures the dynamic feature implied by the variation of social interaction by multi-head self-attention.

4) Intention-Specific Feature Fusion (IFF) Module: The actual trajectory of a vehicle in real traffic scenes is often uncertain due to the complexity and diversity of possible driving maneuvers. Noticing that for multi-lane freeway, the type of driver's maneuver is generally limited, we distinguish lanes keeping (LK), changing lanes to left (CLL) and right (CLR) as three lateral maneuver classes and acceleration (ACC), deceleration (DEC), constant (CON) as three longitudinal maneuver classes. To this end, we introduce intention recognition to obtain the probability of different maneuvers which in turn assists the implementation of multi-modal trajectory prediction described in the next module. In this work, a fully connected layer with softmax activation function is utilized to calculate the probability of lateral maneuver classes and longitudinal maneuver classes as follows:

$$\begin{cases} r = \text{MLP}\left(\tilde{h}_T^0, W_{man}\right) \\ h_{la} = \text{MLP}\left(r, W_{la}\right) \\ P\left(la\right) = \text{softmax}\left(h_{la}\right) \\ h_{lo} = \text{MLP}\left(r, W_{lo}\right) \\ P\left(lo\right) = \text{softmax}\left(h_{lo}\right) \end{cases} \tag{13}$$

where $la \in \{LK, CLL, CLR\}$, $lo \in \{ACC, DEC, CON\}$. Intuitively, the feature vectors acquired from DID module at different timestamps contribute unequally with respect to different prediction horizons. For example, the location of a vehicle at timestamp $T+1$ is more relevant to the location at timestamp T rather than timestamp 1 because of the continuity of motion. The predicted trajectory under different maneuvers should be different thus requiring distinct features as input.

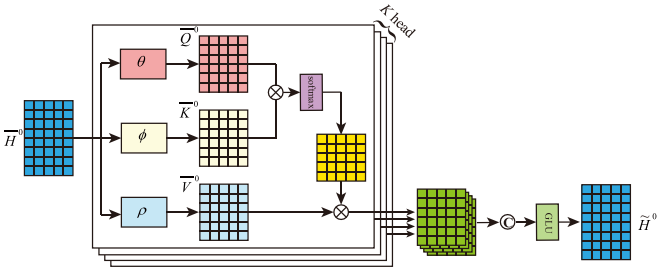


Fig. 4. Illustration of DID module.

However, it is unclear which feature vectors provided by DID module are more relevant than others for correct prediction in terms of a specific horizon. To address this issue, we propose an intention-specific feature fusion (IFF) module to explicitly consider the importance of encoded feature representation by combing features at different historical timestamps in an adaptive manner.

Specifically, given the feature vectors $\tilde{H}^0 = \left[\tilde{h}_1^0, \tilde{h}_2^0, \cdots, \tilde{h}_T^0\right]$ acquired from the DID module, a sequence of weight vector $u_{t'}^{la} = \left[u_{1,t'}^{la}, u_{2,t'}^{la}, \cdots, u_{T,t'}^{la}\right]$ for three lateral maneuvers, $u_{t'}^{lo} = \left[u_{1,t'}^{lo}, u_{2,t'}^{lo}, \cdots, u_{T,t'}^{lo}\right]$ for three longitudinal maneuvers for $t' = T+1, T+2, \cdots, T+F$ is respectively used to combine $\tilde{h}_t^0$ at different timestamps as follows:

$$v_{t'}^{la,lo} = \sum_{t=1}^{T} \left(u_{t,t'}^{la} + u_{t,t'}^{lo}\right) \tilde{h}_t^0 \tag{14}$$

The above weight vectors for lateral and longitudinal maneuvers can be organized in matrix form as $U^{la} = \begin{bmatrix} u_{1,1}^{la} & \cdots & u_{1,F}^{la} \\ \vdots & \ddots & \vdots \\ u_{T,1}^{la} & \cdots & u_{T,F}^{la} \end{bmatrix} \in R^{T \times F}$ and $U^{lo} = \begin{bmatrix} u_{1,1}^{lo} & \cdots & u_{1,F}^{lo} \\ \vdots & \ddots & \vdots \\ u_{T,1}^{lo} & \cdots & u_{T,F}^{lo} \end{bmatrix} \in R^{T \times F}$, where $la \in \{LK, CLL, CLR\}$, $lo \in \{ACC, DEC, CON\}$. The above intention-specific feature fusion is intuitively demonstrated in Fig. 5.

Obviously, the value of $u_{t,t'}^{la}$ and $u_{t,t'}^{lo}$ implies respectively the contribution of feature vector $\tilde{h}_t^0$ when predicting lateral and longitudinal location at timestamp t' for $t' = T+1, T+2, \cdots, T+F$,. More important features tend to have larger weights and vice versa. Notice that the above weight vectors U^{la}, and U^{lo} can be learned from trajectory data in an end-to-end fashion.

5) Multi-Modal Trajectory Prediction (MTP) Module: To estimate the probability distribution of trajectory prediction, we obey the total probability theorem and factorize $P\left(Y|X\right)$ as follows:

$$P\left(Y|X\right) = \sum_{lat,lon} P_\theta\left(Y|X, lat, lon\right) P\left(lat|X\right) P\left(lon|X\right) \tag{15}$$

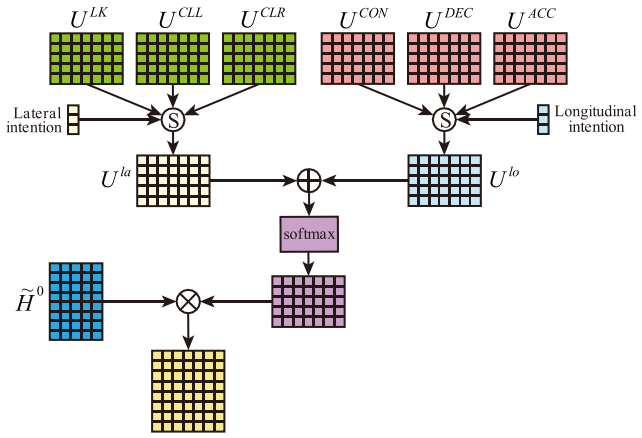


Fig. 5. Intention-specific feature fusion.

where $\theta = \left[\theta_{T+1}, \theta_{T+2}, \cdots, \theta_{T+F}\right]$ denotes the parameters of bivariate Gaussian distribution of the target vehicle position at each timestamp in the future. Each $\theta_{t'} = \left\{\mu_{t',x}, \mu_{t',y}, \sigma_{t',x}, \sigma_{t',y}, \rho_{t'}\right\}$ for $T+1 \le t' \le T+F$ consists of the mean, variance, and correlation coefficient of the vehicle x- and y-position thus reflecting the uncertainty of the prediction. In order to predict the probability distribution of the future trajectory based on different maneuvers, the intention-specific feature $v_{t'}^{la,lo}$ and the prediction probability of maneuver classes are concatenated and then fed into a fully connected layer with weight W_{fc} as follows:

$$e_{t'}^{la,lo} = \mathrm{MLP}\left(v_{t'}^{la,lo} \oplus P\left(la\right) \oplus P\left(lo\right), W_{fc}\right) \tag{16}$$

Additionally, considering the temporal continuity of predicted trajectory, the encoded representation $e_{t'}^{la,lo}$ at each timestamp is sequentially fed into a decoder LSTM which outputs the required predictive distribution parameters $\theta_{t'}$ for vehicle motion, that is

$$h_{t'} = \mathrm{LSTM}\left(e_{t'}^{la,lo}, h_{t'-1}, W_{de-LSTM}\right) \tag{17}$$

$$\theta_{t'} = \mathrm{MLP}\left(h_{t'}, W_{pre}\right) \tag{18}$$

where $W_{de-LSTM}$ denotes the weight in LSTM cell, W_{pre} is the weight used to convert hidden state $h_{t'}$ into the parameters of bivariate Gaussian distribution.

C. Implementation Details

Similar to CS-LSTM [12], we adopt a two-stage training strategy since it can facilitate network parameter learning, especially in the initial stage. Specifically, in the first four epochs, the following Mean Square Error (MSE) loss is minimized due to its simplicity in optimization

$$MSE = \sum_{t'=T+1}^{T+F} \left\{\left(\mu_{t',x} - x_{t'}\right)^2 + \left(\mu_{t',y} - y_{t'}\right)^2\right\} \tag{19}$$

where $(x_{t'}, y_{t'})$ is the ground truth position of the target vehicle at timestamp t', $\left(\mu_{t',x}, \mu_{t',y}\right)$ is the position predicted by the model. Then, we switch to the following negative log likelihood (NLL) loss to further train the network

$$NLL = -\log\left(\sum_{lat,lon} P_\theta\left(Y|X, lat, lon\right) P\left(lat|X\right) P\left(lon|X\right)\right) \tag{20}$$

For each training sample, only one maneuver class is assigned and therefore we choose to optimize the following NLL instead

$$\begin{aligned} NLL &= -\log\left(P_\theta\left(Y|X, lat, lon\right) P\left(lat|X\right) P\left(lon|X\right)\right) \\ &= NLL_{pred} + NLL_{lat} + NLL_{lon} \end{aligned} \tag{21}$$

where $NLL_{pred} = -\log\left(P_\theta\left(Y|X, lat, lon\right)\right)$, $NLL_{lat} = -\log\left(P\left(lat|X\right)\right)$, $NLL_{lon} = -\log\left(P\left(lon|X\right)\right)$. To minimize the loss of maneuver recognition related terms, i.e., NLL_{lat} and NLL_{lon}, we adopt the following cross entropy loss

$$NLL_{lat} = -\sum_{lat \in \{LK, CLL, CLR\}} y_{lat} \log P\left(lat|X\right) \tag{22}$$

$$NLL_{lon} = -\sum_{lon \in \{ACC, DEC, CON\}} y_{lon} \log P\left(lon|X\right) \tag{23}$$

where y_{lat} and y_{lon} denote the ground truth lateral and longitudinal maneuver, respectively. For the trajectory related term, according to bivariate Gaussian distribution, we have

$$\begin{aligned} NLL_{traj} = \sum_{t'=T+1}^{T+F} \Bigg\{ &\log\left(2\pi\sigma_{t',x}\sigma_{t',y}\sqrt{1-\rho_{t'}^2}\right) \\ &+ \frac{1}{2\left(1-\rho_{t'}^2\right)}\left(\frac{\left(\mu_{t',x} - x_{t'}\right)^2}{\sigma_{t',x}^2}\right. \\ &\left. - \frac{\left(\mu_{t',x} - x_{t'}\right)\left(\mu_{t',y} - y_{t'}\right)}{\sigma_{t',x}\sigma_{t',y}} + \frac{\left(\mu_{t',y} - y_{t'}\right)^2}{\sigma_{t',y}^2}\right)\Bigg\} \end{aligned} \tag{24}$$

The PyTorch deep learning framework [49] is used to implement the proposed network. Our model is trained with an Nvidia GTX 1080Ti GPU. The Adam optimizer [50] with an initial learning rate 0.0005 is used to train the network in an end-to-end fashion.

D. Working Pipeline of Network Architecture

First, the raw input trajectory is projected to extract more features by multi-layer perception (MLP), and the resulting feature sequence is fed into a long short-term memory network (LSTM) such that temporal correlation can be encoded in terms of individual motion pattern. Next, the social interaction features at each timestamp are extracted by multi-head spatial attention to reflect the social influence of surrounding vehicles on the target vehicle. Subsequently, the potential dynamic dependency between social interaction across different timestamps is further captured by multi-head temporal self-attention which reflects the varying interactive pattern. Then, the lateral and longitudinal maneuvers are recognized and six combination matrices associated with different maneuvers are learned to fuse the features from different timestamps

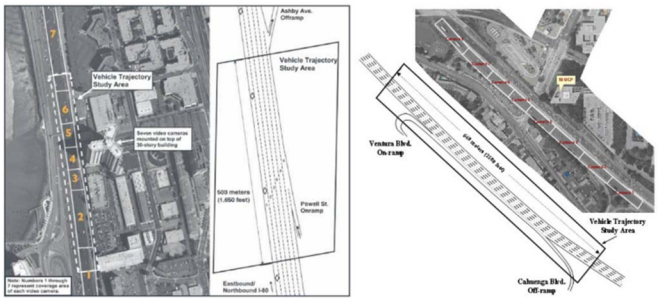


Fig. 6. NGSIM data collection area (a) I-80 (b) I-101 [51].

Fig. 7. Highway drone dataset HighD [53].

based on the predicted maneuver. Finally, another MLP and LSTM-based decoder are utilized to generate multiple possible trajectories.

IV. Experiment Results and Analysis

In this section, we carry out experiments on publicly available datasets to evaluate the proposed STDAN model. In addition, some state-of-the-art algorithms are also included for a comprehensive comparison.

A. Experimental Dataset and Setting

The first dataset used in the experiments is the Next Generation Simulation (NGSIM) [51], [52] consisting of detailed vehicle trajectory information such as vehicle's coordinates, velocity, etc., on eastbound I-80 in the San Francisco Bay area and southbound US 101 in Los Angeles. This dataset was collected by the U.S. Department of Transportation in the year of 2015. This dataset consists of real highway driving scenarios recorded by multiple overhead cameras at 10Hz. The second dataset is the HighD [53] which is collected at a segment of about 420m of two-way roads around Cologne in German from drone video recordings at 25 Hz in the year of 2017 and 2018. It consists of 110 000 vehicles including cars and trucks and a total driven distance of 45 000 km. An illustration of NGSIM and HighD is shown in Fig. 6 and 7, respectively.

B. Comparison Models

- **Social-LSTM (S-LSTM)** [30]: a model in which a shared LSTM is used to encode the raw trajectory data for each vehicle and then the extracted features of different vehicles are aggregated by the social pooling layer.
- **Convolutional Social-LSTM (CS-LSTM)** [12]: Different from S-LSTM, this model captures the social interaction by stacking convolutional and pooling layers and takes into account the multi-modality based on the predicted maneuver.
- **Non-local Social Pooling (NLS-LSTM)** [16]: an LSTM based encoder-decoder structure where the social interaction is captured by combining both local and non-local operations. However, this model lacks multi-modal prediction.
- **Social GAN (S-GAN)** [44]: the model combines a recurrent sequence-to-sequence model and a generative adversarial network to aggregate information across different agents such that multiple socially plausible futures can be generated.
- **Planning-informed prediction (PiP)** [54]: this model couples trajectory prediction as well as the planning of the target vehicle by conditioning on multiple candidate trajectories of the target vehicle.
- **Dual Learning Model (DLM)** [55]: this model embeds an occupancy map and traffic scene risk map into the LSTM-based encoder and decoder structure to learn about the interaction between vehicles and associated risk.
- **Spatial-temporal dynamic attention network (STDAN):** this is the model proposed in this paper.

C. Evaluation Metric

We use the Root Mean Square Error (RMSE) to evaluate the prediction performance of different methods. Specifically, this metric measures the difference between the predicted position and the true position at different timestamps using

$$RMSE=\sqrt{\frac{1}{LF}\sum_{l=1}^{L}\sum_{t'=T+1}^{T+F}\left\{\left(\mu_{t',x}^{l}-x_{t'}^{l}\right)^{2}+\left(\mu_{t',y}^{l}-y_{t'}^{l}\right)^{2}\right\}} \tag{25}$$

where $\left(\mu_{t',x}^{l},\mu_{t',y}^{l}\right)$ and $\left(x_{t'}^{l},y_{t'}^{l}\right)$ are respectively the predicted position and the ground truth position of the target vehicle of the l-th test sample at prediction timestamp t', L denotes the total number of test samples. F is the prediction horizon and in this experiment, F is changed from 1s to 5s. Notice that our model can generate multiple trajectories. For multi-modal prediction, only the predicted trajectory with the highest probability will be used to calculate the RMSE. Otherwise, our model will produce a single trajectory which is used to compute the above metric.

D. Quantitative Results

The quantitative experimental results on NGSIM and HighD datasets are summarized in Table I and shown in Fig. 8. The longitudinal and lateral prediction error of STDAN on NGSIM and HighD datasets is further shown in Fig. 9. As we can see from the results, the involved algorithms can capture the interaction between vehicles by using different strategies and therefore leading to small prediction error. It consolidates that considering surrounding vehicles is a key factor to perform

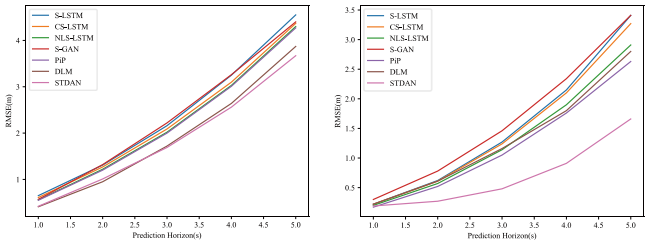


Fig. 8. Variation of prediction error obtained by different methods on NGSIM (left) and HighD (right) datasets.

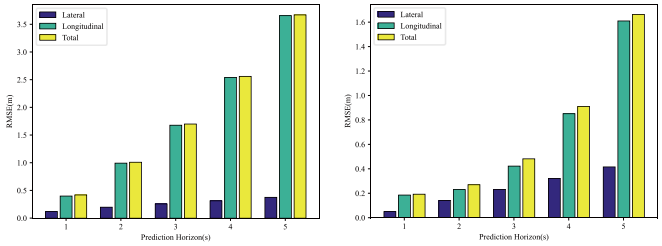


Fig. 9. Longitudinal and lateral prediction error of STDAN on NGSIM (left) and HighD (right) datasets.

trajectory prediction in common traffic sense where more vehicles share the same drivable area. For NGSIM data, our proposed STDAN is slightly lower than the state-of-the-art model for short-term prediction (1-3 seconds), however, works better than all competitors in terms of long-term prediction (3-5 seconds). For HighD data, STDAN outperforms all other competing algorithms, irrespective of the prediction horizons. Moreover, the improvement is enlarged with the increase of horizon. This result is reasonable since, intuitively speaking, the short-term prediction is closely related to the most recent vehicle dynamics while the long-term prediction is more relevant to behavior information, such as intention and driving preference. Moreover, short-term prediction can already achieve a good performance and leaves limited space for improvement. Finally, it can be observed that the prediction error on HighD dataset is significantly smaller than NGSIM dataset for each algorithm. This phenomenon has also been observed by existing studies. It is probably because the trajectory of HighD dataset such as the location, velocity, etc., is more precise than that of NGSIM dataset. Moreover, the sample size of the two datasets is also different: the samples in HighD dataset are about 12 times as many as NGSIM dataset.

In Table II, we report the inference cost of S-LSTM, CS-LSTM, S-GAN, PiP, and our proposed model on NGSIM dataset in terms of both time (ms) and MFLOPS because these models have released source code publicly. As can be seen, the inference speed of our model is fast. It should be noticed that inference cost highly depends on the programming tricks and computation device. Even for the same model, the cost may change significantly due to different implementations.

E. Prediction Visualization

In order to gain qualitative insight, we visualize the prediction results to investigate the performance of our proposed model in different driving scenarios. First, a visualization

TABLE I

PREDICTION ERROR OBTAINED BY DIFFERENT MODELS

Dataset	Horizon (s)	S-LSTM	CS-LSTM	NLS-LSTM	S-GAN	PiP	DLM	STDAN
NGSIM	1	0.65	0.61	0.56	0.57	0.55	**0.41**	0.42
	2	1.31	1.27	1.22	1.32	1.18	**0.95**	1.01
	3	2.16	2.09	2.02	2.22	1.94	1.72	**1.69**
	4	3.25	3.10	3.03	3.26	2.88	2.64	**2.56**
	5	4.55	4.37	4.30	4.40	4.04	3.87	**3.67**
HighD	1	0.22	0.22	0.20	0.30	0.17	0.22	**0.19**
	2	0.62	0.61	0.57	0.78	0.52	0.61	**0.27**
	3	1.27	1.24	1.14	1.46	1.05	1.16	**0.48**
	4	2.15	2.10	1.90	2.34	1.76	1.80	**0.91**
	5	3.41	3.27	2.91	3.41	2.63	2.80	**1.66**

TABLE II

INFERENCE COST OF DIFFERENT MODELS

Metric	S-LSTM	CS-LSTM	S-GAN	PiP	STDAN
Time(ms)	3.022	0.021	1.574	0.315	0.061
MFLOPs	15.898	14.106	57.228	136.250	43.022

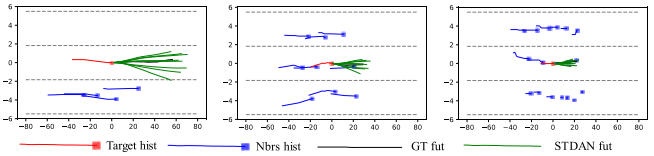


Fig. 10. Examples of multi-modal trajectory prediction.

of multi-modal trajectory prediction under different traffic densities is qualitatively illustrated in Fig. 10. It not only describes the historical trajectory of both target and neighboring vehicles but also plots the multi-modal prediction as well as the ground truth trajectory of the target vehicle. As we can see from Fig. 10, the multi-modal prediction favorably covers the ground truth despite different numbers of surrounding vehicles. Additionally, an interesting phenomenon is that with the decrease of motion speed of the target vehicle (indicated by the shorter historical trajectory), the predicted trajectory gets shorter correspondingly because of slower motion speed.

In Fig. 11, we show some typical prediction results including lane keeping, lane changing to left, and lane changing to right. It also considers different traffic densities, such as light traffic, moderate traffic, and heavy traffic. As we can see from these figures, our proposed model can effectively predict the future trajectory of the target vehicle in different traffic scenes. In summary, these results demonstrate that the reasonable future trajectory is successfully inferred from our model since the sublet and sophisticated temporal dependency and social interaction can be effectively captured by the proposed hierarchical model structure.

Next, we show in Fig. 12 the combination matrix U^{la} and U^{lo} in the proposed Intention-specific Feature Fusion Module. As we can see, the temporal and social features extracted at different timestamps have distinct influences depending on the prediction horizon. Therefore, utilizing different combination matrices to distinguish different driving intentions contributes to the discovery of some sublet but important patterns when performing prediction.

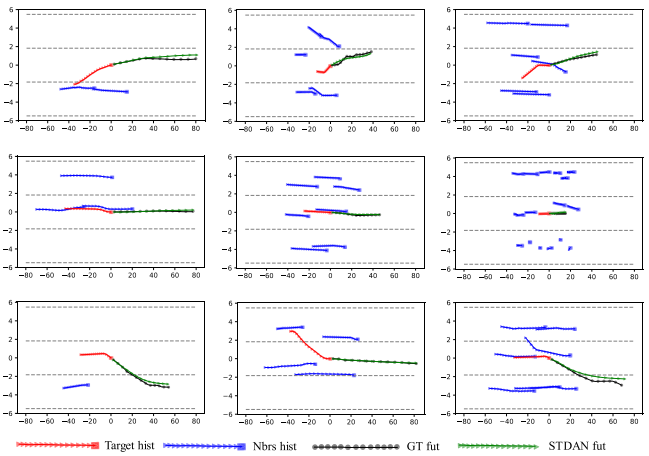


Fig. 11. Illustration of prediction results in different traffic scenes. Top row: lane changing to left, middle row: lane keeping, bottom row: lane changing to right. Left column: light traffic, Middle column: moderate traffic, Right column: heavy traffic.

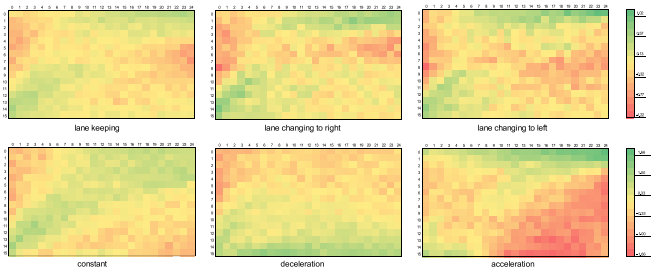


Fig. 12. Illustration of combination matrix learned in IFF module. Top row: longitudinal maneuver, Bottom row: lateral maneuver.

F. Ablation Study

Some ablative experiments are carried out to provide more insights into the behavior of our model especially the impact of different components on the prediction performance by disabling the corresponding component from the entire STDAN. In particular, we consider the following four models, each of which removes a specific component.

STDAN(-ME): remove the Motion Encoder Module and directly feed the raw input trajectory (through a fully connected layer) into the Social Interaction Module. As a result, the temporal correlation information of motion state is ignored.

STDAN(-SI): remove the Social Interaction Module and thus only tackle with the target vehicle without considering any influence of the surrounding vehicles on the target vehicle.

STDAN(-DID): remove the Dynamic Interaction Dependency Module and thus the temporal dependency between social interaction across different timestamps is discarded.

STDAN(-IFF): remove the Intention-specific Feature Fusion Module which ignores the feature combination across different timestamps and only uses the feature obtained at the last timestamp. By doing so, the historical information at previous timestamps cannot be fully exploited when performing intention recognition and trajectory prediction.

The experimental results of the ablation study are reported in Table III. As we can see, after removing different components from STDAN, the obtained model generally suffers from prediction performance degeneration with different

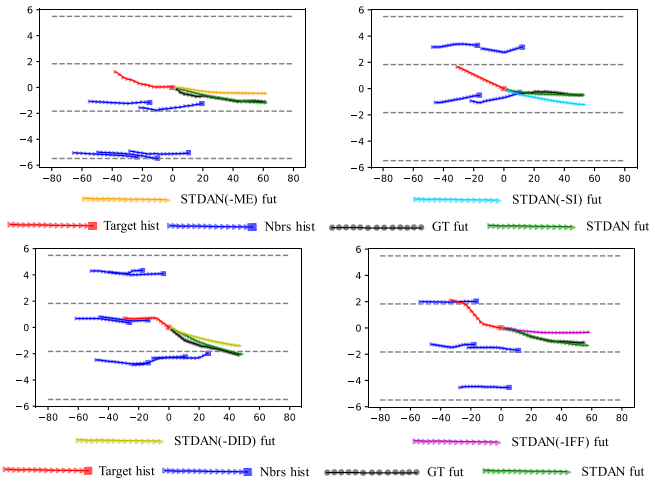


Fig. 13. Illustration of prediction results of different models.

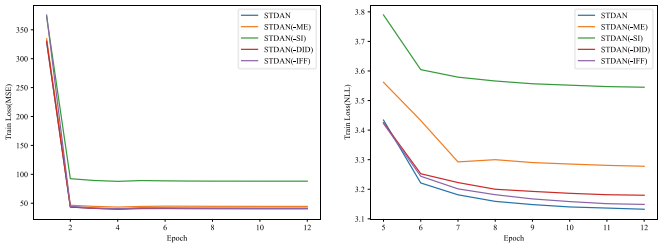


Fig. 14. Variation of training loss in terms of MSE (left) and NLL (right).

degrees. Some interesting observations are discussed as follows. Without the ME module, the overall performance for either short-term or long-term prediction is reduced thus implying LSTM is beneficial for the extraction of temporal information from the low-level motion state of vehicles. When removing SI module, the prediction performance is poor on all predicted horizons. More importantly, with the increase of prediction horizon, the degree of degradation increases remarkably. Therefore, we think that social interaction plays an indispensable role if long-term prediction is required. When removing DID module or IFF module, the prediction error increases with respect to full STDAN. Besides the above prediction error, some quantitative illustrations of the four ablative models are shown in Fig. 13. Some interesting scenes can be identified. For instance, the predicted trajectory of STDAN(-SI) model still follows the historical trajectory and thus may lead to risk collisions with neighboring vehicles because the social interaction module has been removed. In contrast, the full STDAN model can produce a consistent result which changes the driving direction.

In order to analyze the convergence obtained by different ablation models, the variation of loss in terms of both MSE and NLL on the training data is shown in Fig. 14 respectively. As stated above, a two-stage training strategy is adopted where we start training the model with MSE loss (19) in the first four epochs and then switch to NLL loss (24). As shown in Fig. 14, the full STDAN model is able to produce smaller training loss in comparison with all the ablation models which indicates better adaption to the data is achieved due to the elaborated model structure.

TABLE III
PREDICTION ERROR OF ABLATION MODELS

Model	1s	2s	3s	4s	5s
STDAN(-ME)	0.44	1.06	1.79	2.72	3.90
STDAN(-SI)	0.46	1.29	2.42	3.84	5.52
STDAN(-DID)	0.43	1.04	1.75	2.67	3.82
STDAN(-IFF)	0.42	1.03	1.74	2.65	3.81
STDAN	0.42	1.01	1.69	2.56	3.67

TABLE IV
INFERENCE COST OF ABLATION MODELS

Model	Time (ms)	MFLOPs
STDAN(-ME)	0.053	35.879
STDAN(-SI)	0.035	4.770
STDAN(-DID)	0.049	41.056
STDAN(-IFF)	0.056	42.295
STDAN	0.061	43.022

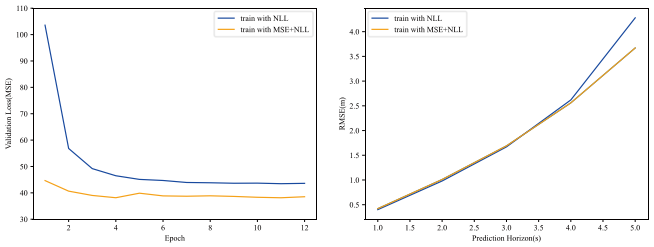


Fig. 15. Variation of validation loss (left) and prediction error (right).

Finally, the inference cost comparison of the above ablation models during the test phase is shown in Table IV. It can be observed that the inference time is in milliseconds for all models, thus enabling the real-time trajectory prediction. Among these ablation models, STDAN(-SI) has the fastest inference speed due to the complete ignorance of social interaction. However, as discussed above, this model suffers from poor prediction performance. The full STDAN is slightly slower than the other ablation models since it captures complex temporal and social information through a hierarchical structure, which leads to significantly improved prediction performance.

G. Discussion

To train the network, we adopt a two-stage strategy where MSE loss is used in the first four epochs and then NLL loss is applied. The influence of MSE loss is experimentally studied and the results are shown in Fig. 15.

As we can see from these results, compared with the model trained with NLL loss, the model trained with MSE plus NLL loss can lead to lower validation loss and smaller prediction error for horizon 4s and 5s. For the possible reason, we conjecture that NLL loss is more complex and thus relatively difficult to optimize with respect to MSE loss. Therefore, if the network is trained with MSE loss in the initial stage, it possibly finds a good solution and thus facilitates the following optimization with NLL loss.

Our model is originally designed for the prediction of the target vehicle by leveraging the information from the surrounding vehicles. If we want to obtain the trajectories of all vehicles in the same scene, the prediction can be efficiently executed

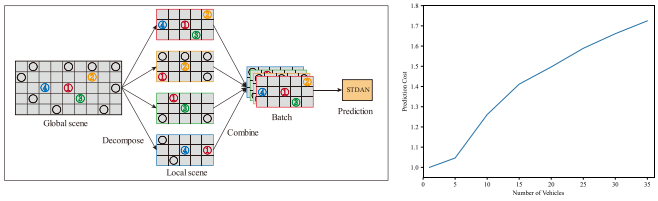


Fig. 16. Multi-vehicle prediction scheme (left) and prediction cost (right).

especially considering that modern computation devices such as GPU are generally equipped with parallel processing power. For example, we can first decompose the global scene into multiple local scenes, each of which centers on a specific vehicle. Then, all those local scenes can be combined into a batch and fed into the prediction model, as shown in Fig. 16 (left). Taking a traffic scene with 35 vehicles for example. Let the cost for predicting one vehicle be 1, the variation of computation cost when increasing the number of vehicles to be predicted is shown in Fig. 16 (right). As can be seen, the prediction cost only grows slowly and thus our model is applicable to multi-vehicle prediction.

V. CONCLUSION

In this paper, we propose a novel vehicle trajectory prediction model by hierarchically extracting features from raw trajectory data of the target and the surrounding vehicles. We explicitly take account of three different information in a layer-by-layer manner, such as motion state, social interaction, as well as the temporal correlation between interactions. Moreover, the whole historical timestamps are utilized so that abundant features can be leveraged for prediction. At the same time, in order to address the inherent multi-modal characteristics of trajectory prediction, an interesting feature fusion mechanism dedicated to different intentions is further developed to capture those sublet but essential features, therefore successfully generating diverse and accurate future trajectories. Finally, extensive quantitative and qualitative experiments on publicly available datasets verify that our proposed model is competitive to other state-of-the-art methods in terms of short-term prediction, however, works much better when conducting long-term prediction. In the future work, extending our model to pedestrian trajectory prediction [56], [57] is interesting since the pedestrian motion analysis is another important application for trajectory analysis. Moreover, we notice that our spatial-temporal feature learning model resembles the "Divided Space-Time Attention" (T+S) architecture proposed in [58]. Therefore, an interesting extension is to fuse social and temporal information jointly following the "Joint Space-Time" attention architecture [58]. In addition, the application of our prediction model to point sequence compression [59] is also a promising direction.

REFERENCES

[1] J. Zhang and K. B. Letaief, "Mobile edge intelligence and computing for the Internet of Vehicles," *Proc. IEEE*, vol. 108, no. 2, pp. 246–261, Feb. 2020.

[2] J. Liu, Y. Luo, H. Xiong, T. Wang, H. Huang, and Z. Zhong, "An integrated approach to probabilistic vehicle trajectory prediction via driver characteristic and intention estimation," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Auckland, New Zealand, Oct. 2019, pp. 3526–3532.

[3] W. Lin *et al.*, "A tube-and-droplet-based approach for representing and analyzing motion trajectories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1489–1503, Aug. 2016.

[4] A. Sarkar and K. Czamecki, "A behavior driven approach for sampling rare event situations for autonomous vehicles," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Macao, China, Nov. 2019, pp. 6407–6414.

[5] S. A. Goli, B. H. Far, and A. O. Fapojuwo, "Vehicle trajectory prediction with Gaussian process regression in connected vehicle environment," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 550–555.

[6] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 935–942.

[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.*, Colorado Springs, CO, USA, Jun. 2011, pp. 1345–1352.

[8] R. Chandra *et al.*, "Forecasting trajectory and behavior of road-agents using spectral clustering in graph-LSTMs," *IEEE Robot. Autom. Lett.*, vol. 5, no. 3, pp. 4882–4890, Jul. 2020.

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

[10] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*.

[11] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[12] N. Deo and M. M. Trivedi, "Convolutional social pooling for vehicle trajectory prediction," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 1468–1476.

[13] L. Shi *et al.*, "SGCN: Sparse graph convolution network for pedestrian trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 8994–9003.

[14] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Montgomery, AL, USA, Dec. 2017, pp. 5998–6008.

[15] Y. Liu, J. Zhang, L. Fang, Q. Jiang, and B. Zhou, "Multimodal motion prediction with stacked transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7577–7586.

[16] K. Messaoud, I. Yahiaoui, A. Verroust-Blondet, and F. Nashashibi, "Non-local social pooling for vehicle trajectory prediction," in *Proc. IEEE Intell. Vehicles Symp.*, Paris, France, Jun. 2019, pp. 975–980.

[17] K. Messaoud, I. Yahiaoui, A. Verroust-Blondet, and F. Nashashibi, "Attention based vehicle trajectory prediction," *IEEE Trans. Intell. Vehicles*, vol. 6, no. 1, pp. 175–185, May 2020.

[18] Y. Yang, Y. Zhuang, and Y. Pan, "Multiple knowledge representation for big data artificial intelligence: Framework, applications, and case studies," *Frontiers Inf. Technol. Electron. Eng.*, vol. 22, no. 12, pp. 1551–1558, 2021.

[19] Y. Pan, "Multiple knowledge representation of artificial intelligence," *Engineering*, vol. 6, no. 3, pp. 216–217, 2020.

[20] H. Xue, D. Q. Huynh, and M. Reynolds, "SS-LSTM: A hierarchical LSTM model for pedestrian trajectory prediction," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Lake Tahoe, NV, USA, Mar. 2018, pp. 1186–1194.

[21] A. Syed and B. T. Morris, "SSeg-LSTM: Semantic scene segmentation for trajectory prediction," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Paris, France, Jun. 2019, pp. 2504–2509.

[22] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 683–700.

[23] D. Helbing and P. Molnár, "Social force model for pedestrian dynamics," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 51, no. 5, p. 4282, 1995.

[24] H. Kim, D. Kim, G. Kim, J. Cho, and K. Huh, "Multi-head attention based probabilistic vehicle trajectory prediction," in *Proc. IEEE Intell. Vehicles Symp.*, Las Vegas, NV, USA, Oct. 2020, pp. 1720–1725.

[25] B. Kim, C. M. Kang, J. Kim, S. H. Lee, C. C. Chung, and J. W. Choi, "Probabilistic vehicle trajectory prediction over occupancy grid map via recurrent neural network," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst. (ITSC)*, Kanagawa, Japan, Oct. 2017, pp. 399–404.

[26] H. Cui *et al.*, "Multimodal trajectory predictions for autonomous driving using deep convolutional networks," in *Proc. Int. Conf. Robot. Automat.*, Montreal, QC, Canada, May 2019, pp. 2090–2096.

[27] R. Quan, L. Zhu, Y. Wu, and Y. Yang, "Holistic LSTM for pedestrian trajectory prediction," *IEEE Trans. Image Process.*, vol. 30, pp. 3229–3239, 2021.

[28] X. Li, X. Ying, and M. C. Chuah, "GRIP: Graph-based interaction-aware trajectory prediction," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Auckland, New Zealand, Oct. 2019, pp. 3960–3966.

[29] N. Nikhil and B. T. Morris, "Convolutional neural network for trajectory prediction," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018.

[30] 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. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 961–971.

[31] H. Zhang, Y. Wang, J. Liu, C. Li, T. Ma, and C. Yin, "A multi-modal states based vehicle descriptor and dilated convolutional social pooling for vehicle trajectory prediction," 2020, *arXiv:2003.03480*.

[32] T. Zhao *et al.*, "Multi-agent tensor fusion for contextual trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 12126–12134.

[33] H. Zhou, D. Ren, H. Xia, M. Fan, X. Yang, and H. Huang, "AST-GNN: An attention-based spatio-temporal graph neural network for interaction-aware pedestrian trajectory prediction," *Neurocomputing*, vol. 445, pp. 298–308, Jul. 2021.

[34] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang, "STGAT: Modeling spatial-temporal interactions for human trajectory prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Seoul, South Korea, Nov. 2019, pp. 6272–6281.

[35] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-STGCNN: A social spatio-temporal graph convolutional neural network for human trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 2020, pp. 14424–14432.

[36] C. Tang and R. Salakhutdinov, "Multiple futures prediction," in *Proc. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2019, pp. 15424–15434.

[37] J. Liang, L. Jiang, K. Murphy, T. Yu, and A. Hauptmann, "The garden of forking paths: Towards multi-future trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA Jun. 2020, pp. 10508–10518.

[38] F. Marchetti, F. Becattini, L. Seidenari, and A. Del Bimbo, "Multiple trajectory prediction of moving agents with memory augmented networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 10, 2020, doi: 10.1109/TPAMI.2020.3008558.

[39] N. Rhinehart, R. Mcallister, K. Kitani, and S. Levine, "PRECOG: Prediction conditioned on goals in visual multi-agent settings," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Seoul, South Korea, Nov. 2019, pp. 2821–2830.

[40] Y. Chai, B. Sapp, M. Bansal, and D. Anguelov, "MultiPath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction," 2019, *arXiv:1910.05449*.

[41] C. Luo, L. Sun, D. Dabiri, and A. Yuille, "Probabilistic multi-modal trajectory prediction with lane attention for autonomous vehicles," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Las Vegas, NV, USA, Oct. 2020, pp. 2370–2376.

[42] X. Feng, Z. Cen, J. Hu, and Y. Zhang, "Vehicle trajectory prediction using intention-based conditional variational autoencoder," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Auckland, New Zealand, Oct. 2019, pp. 3514–3519.

[43] Y. Wang, S. Zhao, R. Zhang, X. Cheng, and L. Yang, "Multi-vehicle collaborative learning for trajectory prediction with spatio-temporal tensor fusion," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 236–248, Jan. 2022.

[44] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 2255–2264.

[45] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2015, *arXiv:1511.07289*.

[46] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. 34th Int. Conf. Mach. Learn.*, Sydney, NSW, Australia, Aug. 2017, pp. 933–941.

[47] J. Lei Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Dec. 2016, pp. 770–778.

[49] A. Paszke *et al.*, "Automatic differentiation in Pytorch," in *Proc. NIPS 2017 Autodiff Workshop*, Long Beach, CA, USA, Oct. 2017

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

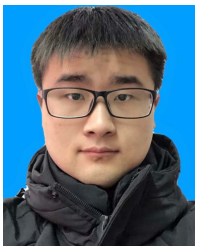
[51] J. Colyar and J. Halkias, "U.S. Highway 101 dataset," Federal Highway Admin., Tech. Rep. FHWA-HRT-07-030, 2007, pp. 27–69.
[52] J. Colyar and J. Halkias, "U.S. Highway 80 dataset," Federal Highway Admin., Tech. Rep. FHWA-HRT-06-137, 2006.
[53] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, "The highd dataset: A drone dataset of naturalistic vehicle trajectories on German highways for validation of highly automated driving systems," in *Proc. Int. Conf. Intell. Transp. Syst.*, Honolulu, HI, USA, Nov. 2018, pp. 2118–2125.
[54] H. Song, W. Ding, Y. Chen, S. Shen, M. Y. Wang, and Q. Chen, "PiP: Planning-informed trajectory prediction for autonomous driving," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 598–614.
[55] M. Khakzar, A. Rakotonirainy, A. Bond, and S. G. Dehkordi, "A dual learning model for vehicle trajectory prediction," *IEEE Access*, vol. 8, pp. 21897–21908, 2020.
[56] W. Lin *et al.*, "Human in events: A large-scale benchmark for human-centric video analysis in complex events," 2020, *arXiv:2005.04490*.
[57] Y. Cai *et al.*, "Pedestrian motion trajectory prediction in intelligent driving from far shot first-person perspective video," *IEEE Trans. Intell. Transp. Syst.*, early access, Jan. 28, 2021, doi: 10.1109/TITS.2021.3052908.
[58] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" 2021, vol. 2, no. 4, pp. 175–185, *arXiv:2102.05095*.
[59] W. Lin *et al.*, "Key-point sequence lossless compression for intelligent video analysis," *IEEE Multimedia Mag.*, vol. 27, no. 3, pp. 12–22, Jul. 2020.

Xiaobo Chen (Member, IEEE) received the B.S. and M.S. degrees from Jiangsu University in 2004 and 2007, respectively, and the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology, Nanjing, in 2013. From 2015 to 2017, he was a Post-Doctoral Researcher with the University of North Carolina at Chapel Hill. Before July 2021, he was a Professor with the Automotive Engineering Research Institute, Jiangsu University. He is currently a Distinguished Professor with the School of Computer Science and Technology, Shandong Technology and Business University. He is also a Taishan Scholar of Shandong Province of China. He has published more than 100 scientific papers in a variety of international journals and conferences. His research interests include intelligent transportation systems and machine learning.

Huanjia Zhang received the B.S. degree in process equipment and control engineering from Xi'an Shiyou University, Xi'an, China, in 2018. He is currently pursuing the M.S. degree with the School of Computer Science and Technology and the School of Artificial Intelligence, Shandong Technology and Business University. His research interests include trajectory prediction, semantic segmentation, and autonomous driving.

Feng Zhao received the M.S. and Ph.D. degrees in computer science from Xi'dian University, China, in 2004 and 2008, respectively. He is currently a Professor with Shandong Technology and Business University. His research interests are in the areas of pattern recognition and information processing.

Yu Hu received the B.S. degree in traffic engineering from Jilin University, Changchun, China, in 2018. He is currently pursuing the M.S. degree with the Automotive Engineering Research Institute, Jiangsu University. His research interests include vehicle trajectory estimation and vehicle re-identification.



Chenkai Tan received the M.S. degree from the School of Vehicle and Traffic Engineering, Jiangsu University of Technology, Jiangsu, China, in 2021. He is currently pursuing the Ph.D. degree with the Automotive Engineering Research Institute, Jiangsu University, Zhenjiang. His research interests include human trajectory prediction and vehicle trajectory prediction.

Jian Yang (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology (NUST) in 2002. In 2003, he was a Post-Doctoral Researcher with the University of Zaragoza. From 2004 to 2006, he was a Post-Doctoral Fellow with the Biometrics Centre, The Hong Kong Polytechnic University. From 2006 to 2007, he was a Post-Doctoral Fellow with the Department of Computer Science, New Jersey Institute of Technology. From September 2007 to February 2022, he was a Changjiang Distinguished Professor with the School of Computer Science and Engineering, NUST. Since February 2022, he has been a Distinguished Professor with the College of Computer Science, Nankai University. He is the author of more than 200 scientific papers in pattern recognition and computer vision. His papers have been cited more than 26000 times in the Scholar Google. His research interests include pattern recognition, computer vision, and machine learning. He is a fellow of IAPR. Currently, he is an Associate Editor of *Pattern Recognition Letters*, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and *Neurocomputing*.