A Spatial-Temporal Attention Model for Human Trajectory Prediction

Xiaodong Zhao, Yaran Chen, Jin Guo, and Dongbin Zhao, *Fellow, IEEE*

***Abstract*—Human trajectory prediction is essential and promising in many related applications. This is challenging due to the uncertainty of human behaviors, which can be influenced not only by himself, but also by the surrounding environment. Recent works based on long-short term memory (LSTM) models have brought tremendous improvements on the task of trajectory prediction. However, most of them focus on the spatial influence of humans but ignore the temporal influence. In this paper, we propose a novel spatial-temporal attention (ST-Attention) model, which studies spatial and temporal affinities jointly. Specifically, we introduce an attention mechanism to extract temporal affinity, learning the importance for historical trajectory information at different time instants. To explore spatial affinity, a deep neural network is employed to measure different importance of the neighbors. Experimental results show that our method achieves competitive performance compared with state-of-the-art methods on publicly available datasets.**

***Index Terms*—Attention mechanism, long-short term memory (LSTM), spatial-temporal model, trajectory prediction.**

I. Introduction

HUMAN trajectory prediction is to predict future path according to the history trajectory. The trajectory is represented by a set of sampled consecutive location coordinates. Trajectory prediction is a core building block for autonomous moving platforms, and the prospective applications include autonomous driving [1]–[3], mobile robot navigation [4], assistive technologies [5], and smart video surveillance [6], etc.

When a person is walking in the crowd, the future path is determined by various factors like the intention, the social conventions and the influence of nearby people. For instance, people prefer to walk along the sidewalk rather than crossing the highway. A person is able to adjust his path by estimating the future path of the people around him, and the people do the same thing which in turn affects the target. Human trajectory prediction becomes an extremely challenging problem due to such complex nature of the people. Benefiting from the powerful deep learning [7], [8], human trajectory prediction has gained a significant improvement in the last few years. Yagi *et al.* in [5] present a multi-stream convolution-deconvolution architecture for first-person videos, which verifies pose, scale, and ego-motion cues are useful for the future person localization. Pioneering works by [9], [10] shows that long-short term memory (LSTM) has the capacity to learn general human movements and predict future trajectories.

Although tremendous efforts have been made to address these challenges, there are still two limitations:

1) The historical trajectory information at different time instants has different levels of influence on the target human, which is ignored by most of works. However, it plays an important role on the prediction of the future path. As for the target human, the latest trajectory information usually has a higher level of influence on the future path as shown in Fig. 1(a). As for the neighbors, the trajectory information will have a great impact as long as the distance is close to the target, as shown in Fig. 1(b). Thus, the historical trajectory information at different time instants ought to be given different weights. The attention mechanism is capable of learning different weights according to the importance.

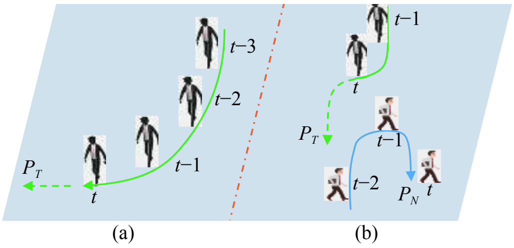

Fig. 1. Illustration of the influences at different time instants. (a) As for the target human (P_T), the trajectory information at time $t-1$ and t may affect future path more compared with that at time $t-2$ and $t-3$. (b) As for the neighbor (P_N), he turns away from P_T at time t. The trajectory information of P_N at time $t-1$ has a greater influence on P_T considering that P_T is not allowed to occupy the position where P_N just lefts.

Manuscript received March 24, 2020; accepted April 13, 2020. This work was supported by the National Key Research and Development Program of China (2018AAA0101005, 2018AAA0102404), the Program of the Huawei Technologies Co. Ltd. (FA2018111061SOW12), the National Natural Science Foundation of China (61773054), and the Youth Research Fund of the State Key Laboratory of Complex Systems Management and Control (20190213). Recommended by Associate Editor Guangjie Han. *(Corresponding author: Yaran Chen and Jin Guo.)*

Citation: X. D. Zhao, Y. R. Chen, J. Guo, and D. B. Zhao, "A spatial-temporal attention model for human trajectory prediction," *IEEE/CAA J. Autom. Sinica*, vol. 7, no. 4, pp. 965–974, Jul. 2020.

X. D. Zhao is with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, and also with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: s20180612@xs.ustb.edu.cn).

Y. R. Chen and D. B. Zhao are with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: chenyaran2013@ia.ac.cn; dongbin.zhao@ia.ac.cn).

J. Guo is with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, and also with the Key Laboratory of Knowledge Automation for Industrial Processes, Ministry of Education, Beijing 100083, China (e-mail: guojin@ustb.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2020.1003228

2) Most of trajectory prediction methods fail to capture the global context among the environment. Some methods capture the global context through an annotation text recording people location coordinates provided by the dataset. However, the text just annotates a few people, so it is not the real global information. A pre-trained detection model [11] can be used to extract all people in the image rather than relying on the annotation text.

In this work, we propose a spatial-temporal attention network to predict human trajectory in the future. We adopt an LSTM called ego encoder to model the ego motion of the target human. We also consider all people in the scene by the pre-trained detection model extracting the positions of neighbors. The positions are fed into a multi-layer perceptron (MLP) to obtain high dimensional features. Then the inner product is used to acquire the weights which measure the importance of neighbors to the target. Further, another LSTM called interaction encoder is followed to model human-human interaction. It is noted that in most existing models, the trajectory information at different time instants gains equal treatment, which is not suitable for the complex trajectory prediction. Inspired by this, we introduce an attention mechanism to obtain the weights, which represent the levels of influence for trajectory information at different time instants. Finally, an LSTM decoder is employed to generate human trajectories for the next few frames.

Our contributions can be summarized as following:

1) We introduce an attention mechanism to automatically learn the weights. They dynamically determine which time instant's trajectory information we should pay more attention to.

2) We utilize a pre-trained detection model [11] to capture global context instead of retrieving local context from the dataset, then an MLP and the inner product are used to weight different neighbors.

3) Based on the above two ideas, a spatial-temporal attention (ST-Attention) model, is proposed to tackle the challenges of trajectory prediction. ST-Attention achieves competitive performance on two benchmark datasets: ETH & UCY [12], [13] and ActEV/VIRAT [14].

II. Related Work

A. Traditional Approaches for Trajectory Prediction

Kalman filter [15], [16] can be deployed to forecast the future trajectory in the case of linear acceleration, which has proven to be an efficient recursive filter. It is capable of estimating the state of a dynamic system from a series of incomplete and noisy measurements, especially in the analysis of time sequences. Williams [17] proposes to use Gaussian processes distribution to estimate motion parameters like the velocity and the angle offset, then a motion pattern of the pedestrian is built. Further, researchers begin to associate the energy with pedestrians. One representative work is the social forces proposed by Helbing and Molnár [18], which transforms the attraction and the exclusion between pedestrians and obstacles into energy to predict the pedestrian trajectory. The attractive force is used to guide the target to the destination, and the repulsive force is used to keep a safe distance and avoid collision. Subsequently, some methods [19] fit the parameters of the energy functions to improve the social forces model.

However, the above methods rely on hand-crafted features. This becomes an obstacle to advance the performance of the trajectory prediction in light that these methods have the ability to capture simple interaction but fail in complex scenarios. In contrast, data-driven methods based on convolutional neural network (CNN) and recurrent neural network (RNN) overcome the above limitations of traditional ones.

B. CNN Models for Trajectory Prediction

CNN [20] has proven to be powerful to extract rich context information, which is salient cue for trajectory prediction task. Behavior-CNN in [21] employs a large receptive field to model the walking behaviors of pedestrians and learn the location information of the scene. Yagi *et al.* [5] develop a deep neural network that utilizes the pose, the location-scale, and the ego-motion cues of the target human, but they forget to consider human-human interaction. Huang *et al.* [22] introduce the spatial matching network and the orientation network. The former generates a reward map representing the reward of every pixel on the scene image, and the latter outputs an estimated facing orientation. However, this method can only consider the static scene but not the dynamic information of pedestrians.

C. RNN Models for Trajectory Prediction

RNN [23] has also proven to be efficient to deal with time sequence tasks. RNN models have shown dominant capability in various domains like neural machine translation [24], speech recognition [25], generating image descriptions [26] and DNA function prediction [27]. Some recent works have attempted to use RNN to forecast trajectory. Social-LSTM in [9] introduces a social pooling layer to learn classic interactions that happen among pedestrians. But this pooling solution fails to capture global context. Besides, social-LSTM predicts the distribution of the trajectory locations instead of directly predicting the locations. This makes training process difficult while sampling process is non-defferentiable. Gupta *et al.* [10] propose Social-GAN combining approaches for trajectory prediction and generative adversarial networks. But the performance has not been improved obviously when sampling only one time during test time. Liang *et al.* [28] present *Next*, an end-to-end learning framework extracting rich visual information to recognize pedestrian behaviors. Furthermore, the focal attention [29] is employed to the framework. It is originally proposed to tackle visual question answering, projecting different features into a low-demensional space. But the focal attention used in *Next* is hard-wired and fails to learn from the data. Xu *et al.* [30] design a crowd interaction deep neural network which considers all pedestrians in the scene as well as their spatial affinity for trajectory prediction. However, they ignore the influence of temporal affinity. In our work we take into account both spatial affinity and temporal affinity.

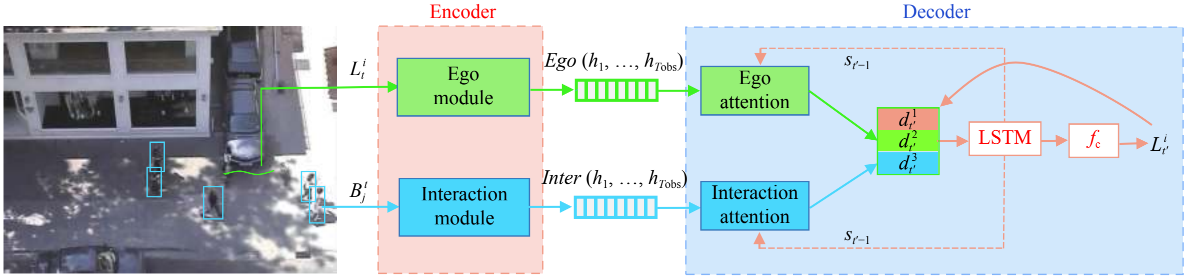

Fig. 2. Overview of proposed ST-Attention. The model utilizes encoder to extract ego feature $Ego(h_1,\ldots,h_{T_{\text{obs}}})$ and interaction feature $Inter(h_1,\ldots,h_{T_{\text{obs}}})$ from L_t^i and B_t^j respectively ($t \in [1, T_{\text{obs}}]$), then the following decoder outputs future path $L_{t'}^i$ ($t' \in [T_{\text{obs}+1}, T_{\text{pred}}]$).

D. Attention Approaches for Trajectory Prediction

Some approaches for trajectory prediction employ the attention mechanism to differentiate the influence of neighbors on the target. Su *et al.* [31] update the LSTM memory cell state with a coherent regularization, which computes the pairwise velocity correlation to weight the dependency between the trajectories. Further, a social-aware LSTM unit [32] is proposed, which incorporates the nearby trajectories to learn a representation of the crowd dynamics. Zhang *et al.* [33] utilize motion gate to highlight the important motion features of neighboring agents. Sadeghian *et al.* [34] apply the soft attention similarly with [26], and emphasis the salient regions of the image and the more relevant agents. However, the above works focus on the spatial influence of the neighboring agents, but ignore the temporal influence of the agents which is also valuable for human trajectory prediction. The attention mechanism in our model connects the decoder state and the temporal encoder state, allowing to give an importance value for each time instant's trajectory state of the neighboring humans and the target human.

E. Pedestrian Re-identification for Trajectory Extraction

With the advancement of the pedestrian re-identification (Re-ID) [35], the same person with different appearances can be identified accurately, which facilitates the extraction of the human trajectory. Köstinger *et al.* [36] consider that the difference among various image features of the same pedestrian conforms to Gaussian distribution, and propose keep-it-simple-and-straightforward metric learning (KISSME). However, KISSME meets the small sample size problem in calculating various classes of covariance matrices, which blocks the improvement of the Re-ID performance. Han *et al.* in [37] verify that virtual samples can alleviate the small sample size problem of KISSME. And the re-extraction process of virtual sample features is eliminated by genetic algorithm, which greatly improves the matching rate of pedestrian Re-ID. Further, KISS+ [38] algorithm is proposed to generate virtual samples by using an orthogonal basis vector, which is very suitable for real-time pedestrian Re-ID in open environment due to its advantages of simplicity, fast execution and easy operation. These works are of great significance to the human trajectory prediction.

III. Method

A person adjusts his trajectory based on the definite destination in mind and the influence of neighbors when he is walking in the crowd. On the one hand, the future trajectory of the target human depends on historical trajectories at different time instants, which we refer to as temporal affinity. On the other hand, the future trajectory hinges on the distances, the velocities and the headings of neighbors, which we refer to as spatial affinity. This idea motivates us to study the trajectory prediction jointly with temporal and spatial affinities. In this section, we present our spatial-temporal attention model tackling the problem.

A. Problem Formulation

We assume that the location coordinates of the target human at different time instants are obtained. Additionally, bounding boxes of all people are extracted through a pre-trained detection model [11]. Suppose that there are n pedestrians, and we denote the location of the ith pedestrian p_i ($i \in [1,n]$) at time t as $L_t^i = (x_t, y_t)$. The observed box of the jth pedestrian p_j ($j \in [1,n]$) at time t is defined as B_t^j. Given the locations of the target human and the bounding boxes of nearby people from time 1 to T_{obs}, our system aims at predicting the locations of the target human from time $T_{\text{obs}+1}$ to T_{pred}. We assume $t \in [1, T_{\text{obs}}]$ and $t' \in [T_{\text{obs}+1}, T_{\text{pred}}]$.

B. Overview

The overall network architecture is illustrated in Fig. 2. Our model employs an encoder-decoder framework. Specifically, the encoder consists of ego module and interaction module, and the decoder includes attention module and prediction module. We feed the locations into ego module to get the ego feature which is used for modeling the motion of the target. At the same time, the observed boxes are fed into interaction module to get the interaction feature which is used for exploring the relationship among neighbors. The above feature vectors are weighted and summed along the temporal dimensions by the attention module. Then the prediction module employs an LSTM to generate future trajectory. In the rest of this section, we will detail the above modules.

C. Ego Module

Ego module aims at exploring the intention of the target human which can be reflected by the motion characteristics such as the velocity, the acceleration and the direction. Due to the powerful ability of addressing sequence data, LSTM is chosen as the ego module architecture. For the pedestrian p_i, we embed the location into a vector e_t. Then the embedding is fed into the *ego encoder*, whose hidden state h_t is computed by

$$\mathrm{e}_t = \phi(L_t^i; W_e) \tag{1}$$

$$h_t = LSTM(h_{t-1}, e_t; W_{en}) \tag{2}$$

where L_t^i represents the location coordinate of p_i at time t and $\phi(\cdot)$ is an embedding function with ReLU non-linearity. We denote embedding weights as W_e and LSTM weights as W_{en}. Finally, we acquire a feature representation $Ego(h_1,\ldots,h_{T_{\mathrm{obs}}})$ with the shape of $T_{\mathrm{obs}} \times d$, where d is the hidden size of the LSTM.

D. Interaction Module

A person changes his trajectory by observing the movements of neighbors. Ego module alone has no capacity to capture the relationship among nearby people. Interaction module is introduced to address this problem as shown in Fig. 3. Some recent works have attempted to extract the motion information from location coordinates while Yagi *et al.* [5] indicate the scale of the human serves as a significant cue to estimate the future trajectory. The observed box reflecting both location and scale is useful for modeling human-human interaction. Besides, Xu *et al.* [30] show that the optical spatial affinity measure can be automatically learned by a nonlinear function and the inner product operation.

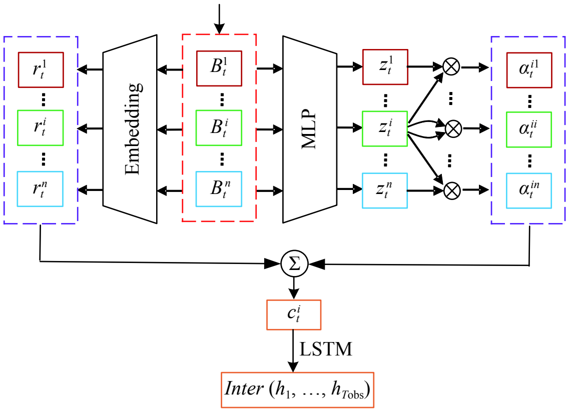

Fig. 3. The framework of the interaction module. Given a sequence of observed boxes, we extract the spatial affinity among nearby people.

Given the observed box B_t^j for pedestrian p_j at time t, we employ a multi-layer perceptron to map the box into the fixed length vector z_t^j, then the spatial affinity can be obtained by

$$\alpha_t^{ij} = \frac{\exp\left(\langle z_t^i, z_t^j\rangle\right)}{\sum\limits_j \exp\left(\langle z_t^i, z_t^j\rangle\right)} \tag{3}$$

where $\langle\cdot,\cdot\rangle$ is the inner product operation, that is, point multiplication of matrices. The spatial affinity of pedestrian p_j to pedestrian p_i at time t is denoted as α_t^{ij}, which is normalized through a softmax function.

Besides, we embed the observed box B_t^j into a fixed-length vector r_t^j, then the influence of nearby people on the target human p_i at each time t, c_t^i, is summed along the spatial dimensions as follows:

$$r_t^j = \phi(B_t^j; W_r) \tag{4}$$

$$c_t^i = \sum_j \alpha_t^{ij} r_t^j \tag{5}$$

where W_r is the embedding weights. Likewise, the interaction encoder is applied to obtain a spatial feature representation $Inter(h_1,\ldots,h_{T_{\mathrm{obs}}})$ of $T_{\mathrm{obs}} \times d$.

E. Attention Module

Human trajectory prediction can be regarded as a sequence to sequence problem, which takes the location sequences as input and then outputs another location sequences. Meanwhile, neural machine translation (NMT) [24] is a matter of considerable concern in sequence to sequence learning. Coincidentally, the models proposed recently for NMT also belong to the family of encoder-decoder. And NMT is often accompanied by an attention mechanism which helps it cope effectively with long input sequences. One representative work is transformer [39], which specializes in learning long-range dependencies and discards RNN due to its sequential computation. And the self-attention weight in transformer is computed by a dot-product function of the query with the corresponding key. This requires that query and key matrices have the same dimension, and the dot-product function has no learnable parameters. In our paper we adopt the soft attention introduced by Bahdanau *et al.* [40] which extends the encoder-decoder model by jointly learning to align and translate with the help of attention. The attention [40] we used is learnable and allows encoder and decoder to have different hidden sizes, which is more flexible and applicable to our algorithm.

In fact, the historical location sequences at different time instants have different influences on the future trajectory. Through the attention mechanism, the decoder decides which part of the input sequences to be paid attention to, which avoids encoding all information from time 1 to T_{obs}.

The ego module and the interaction module both output a feature representation of the shape $T_{\mathrm{obs}} \times d$. The attention mechanism is shown in Fig. 4. We take the output of the ego module $Ego(h_1,\ldots,h_{T_{\mathrm{obs}}})$ as an example to illustrate it. For example, $Ego(h_1,\ldots,h_{T_{\mathrm{obs}}})$ is fed into the attention module, and we aim to predict the trajectory location of the target human at time t'. The k-th sequence of $Ego(h_1,\ldots,h_{T_{\mathrm{obs}}})$ is denoted as h_k, and the hidden state of the LSTM decoder at last time $t'-1$ is denoted as $s_{t'-1}$. Given an encoder state h_k and the decoder state $s_{t'-1}$, the attention score for each pair is calculated by

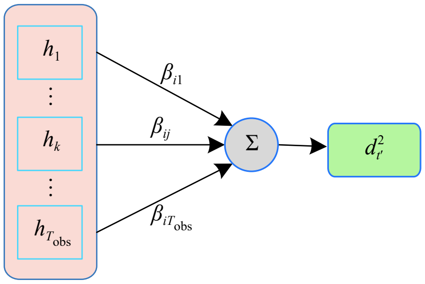

Fig. 4. The graphical illustration of the attention module.

$$score(s_{t'-1}, h_k) = v^T \tanh(W_1 s_{t'-1} + W_2 h_k) \quad (6)$$

where W_1, W_2 and v are learnable parameters. Note that $s_{t'-1}$ is similar to the query matrix and h_k is similar to corresponding key matrix in transformer [39], but the dot-product function is replaced by a learnable function. The attention score reflects dependency relationship between the encoder state h_k and the previous decoder state $s_{t'-1}$. Then all the attention scores subtract the maximum of them to prevent them from growing large in magnitude, which makes training more stable. And the probability $\beta_{t'k}$ is computed as follows:

$$u_{t'k} = score(s_{t'-1}, h_k) - \max_{k=1}^{T_{\text{obs}}} score(s_{t'-1}, h_k) \quad (7)$$

$$\beta_{t'k} = \frac{\exp(u_{t'k})}{\sum_{k=1}^{T_{\text{obs}}} \exp(u_{t'k})} \quad (8)$$

where $\beta_{t'k}$ reflects the importance of the encoder state h_k in determining the next state s'_t and generating next trajectory location at time t'. Then the decoder feature vector $d^2_{t'}$ is given as a weighted sum along the temporal dimensions

$$d^2_{t'} = \sum_{k=1}^{T_{\text{obs}}} \beta_{t'k} h_k \in \mathbb{R}^d. \quad (9)$$

If an input sequence at a certain time instant plays a more important role, it will occupy a larger proportion in the decoder feature vector.

F. Prediction Module

Through attention mechanism, the ego module and the interaction module will produce the d-dimension decoder feature vectors $d^2_{t'}$ and $d^3_{t'}$ respectively. Besides, we embed the xy-coordinate from the last time instant into another decoder feature vector $d^1_{t'}$. The above features are concatenated into a tensor $q_{t'}$ which is fed to the LSTM decoder to get the decoder state $s_{t'}$

$$s_{t'} = LSTM(s_{t'-1}, q_{t'}; W_{de}). \quad (10)$$

We directly predict the location $L^i_{t'}$ of target p_i at time t' from $s_{t'}$ followed by a fully connected layer.

IV. Performance Analysis

In this section, we analyze our model on pedestrian trajectory datasets based on world plane and image plane. Specifically, we evaluate meter values on ETH [12] and UCY [13] datasets, and report pixel values on ActEV/VIRAT dataset [14]. Experimental results demonstrate that our model performs well on both world plane and image plane.

A. Evaluation Metrics

Similarly to prior works [10], [28], we use two metrics to report prediction error:

1) Average Displacement Error (ADE): Average L_2 distance between the ground truth coordinates and the prediction coordinates over all predicted time instants

$$ADE = \frac{\sum_{i=1}^{N} \sum_{t=T_{\text{obs+1}}}^{T_{\text{pred}}} \|L^i_t - \hat{L}^i_t\|_2}{N(T_{\text{pred}} - T_{\text{obs}})}. \quad (11)$$

2) Final Displacement Error (FDE): The L_2 distance between the true points and the prediction points at the final time instant T_{pred}

$$FDE = \frac{\sum_{i=1}^{N} \|L^i_{T_{\text{pred}}} - \hat{L}^i_{T_{\text{pred}}}\|_2}{N}. \quad (12)$$

B. Baseline Methods

We compare the results of our model with following state-of-the-art methods at the same conditions:

1) Linear [10] : A linear regression model whose parameters are determined by minimizing least L_2 distance.

2) S-LSTM [9] : Alahi *et al.* [9] build one LSTM for each person and share the information between the LSTMs through the social pooling layer. At each time instant t during the prediction period, the LSTM hidden-state represents a bivariate Gaussian distribution described by mean μ, standard deviation σ and correlation coefficient ρ. Then the predicted trajectory at time $t+1$ is sampled from the distribution.

3) Next [28] : Liang *et al.* [28] encode a person through rich visual features rather than oversimplifying human as a point. The person behavior module is proposed to capture the visual information, modeling appearance and body movement. And the person interaction module is used to capture other objects information and the surroundings. The future trajectory is predicted by the LSTM with the focal attention [29].

4) LiteNext: We implement a simplified version of *Next* model which just takes into account the person's trajectory and the person-objects. We keep the same in other settings. In this way, the input of LiteNext is the same as ours.

C. Experiments on ETH and UCY

The ETH dataset consists of 750 pedestrians and contains two sets (ETH and HOTEL). The UCY dataset embodies 786 pedestrians and is comprised of three sets (ZARA1, ZARA2 and UNIV). These datasets contain rich real-world scenarios including walking in company, giving way for each other and lingering about, which are full of challenges. The number of tags including frame, pedestrian, group, and obstacle in the datasets is summarized as shown in Table I.

TABLE I
The Number of Tags

Dataset	Frame	Pedestrian	Group	Obstacle
ETH	1448	360	243	44
HOTEL	1168	390	326	25
UNIV	541	434	297	16
ZARA1	866	148	91	34
ZARA2	1052	204	140	34

1) Setup: Following the same experimental setup as [10], we use the leave-one-out strategy, that is, training on 4 sets

TABLE II
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON ETH & UCY DATASETS. WE USE ADE AND FDE TO MEASURE PREDICTION ERROR IN METER VALUES, AND LOWER IS BETTER

Metric	Dataset	Linear	S-LSTM [9]	Next [28]	LiteNext	ST-Attention (ours)
ADE	ETH	1.33	1.09	0.88	0.90	**0.85**
	HOTEL	0.39	0.79	0.36	0.39	**0.32**
	UNIV	0.82	0.67	**0.62**	0.65	0.63
	ZARA1	0.62	0.47	0.42	0.45	**0.42**
	ZARA2	0.77	0.56	0.34	0.36	**0.34**
Average		0.79	0.72	0.52	0.55	**0.51**
FDE	ETH	2.94	2.35	1.98	2.02	**1.85**
	HOTEL	0.72	1.76	0.74	0.80	**0.66**
	UNIV	1.59	1.40	**1.32**	1.37	1.33
	ZARA1	1.21	1.00	**0.90**	0.96	0.91
	ZARA2	1.48	1.17	0.75	0.81	**0.73**
Average		1.59	1.54	1.14	1.19	**1.10**

and testing on the remaining set. Based on the sampling period of 0.4 s, we observe the trajectory for 8 frames (3.2 s) and predict the next 12 frames (4.8 s), namely $T_{\text{obs}} = 8$, $T_{\text{obs}+1} = 9$, and $T_{\text{pred}} = 20$.

2) Implementation Details: In the interaction module, a multi-layer perceptron is employed, which embodies 3 layers. The node sizes in these layers are set to 32, 64, 128 respectively. And the dimension of embedding layer is 128. The LSTM hidden size d is set to 256. A batch size of 64 is used and the epoch number of the training stage is 100. We use Adam optimizer with an initial learning rate of 0.001. To facilitate the training, we clip the gradients at a value of 10. A single NVIDIA GeForce GTX Titan-Xp GPU is used for the training.

3) Quantitative Results: We report the experimental results about ADE and FDE for all methods across the crowd sets in Table II. The linear regressor presents high prediction errors since it has no ability to model curved trajectories unlike other methods employing LSTM networks to overcome this deficiency. Moreover, *Next* model and ours are better than S-LSTM as they consider global human-human interaction, and S-LSTM does not perform as well as expected since it fails to consider global context.

Our ST-Attention achieves state-of-the-art performance on ETH and UCY benchmarks. Throughout the Table II, the evaluation error on single ETH crowd set is much higher than those on other sets. This crowd set contains a lot of pedestrians on the narrow passage and they walk in disorder with different velocities and headings. Compared with LiteNext, the same input information is obtained but ST-Attention performs significantly better, especially on ETH dataset. This verifies the effectiveness of our method. Compared with *Next*, ST-Attention misses two input feature channels and has a lighter network structure. At the same time, ST-Attention is still competitive and achieves powerful results no worse than *Next*. This is because the focal attention [29] used in *Next* is hard-wired and cannot make full use of input features.

Computational time is crucial in order to satisfy real-world applications. For instance, real-time prediction of the pedestrian trajectories in front of the vehicles is necessary in autonomous driving. In Table III, we make a comparison with other models in terms of speed. We can see that S-LSTM has fewer parameters, but the computational time is not as fast as expected. The decrease of speed is because that S-LSTM adopts recursive method to predict future trajectories, which means S-LSTM needs to compute occupancy grids to implement social pooling at each time instant. Compared with *Next*, our method reduces the number of parameters by almost half, since ST-Attention uses fewer input channels. Correspondingly, our method is 2.5x faster than *Next*, taking about 0.02 s on the premise of that L_t^i and B_t^j are obtained. Due to the efficient interaction model, our model is also faster than LiteNext.

TABLE III
SPEED COMPARISON WITH BASELINE MODELS. OUR METHOD WITH FEWER NUMBER OF PARAMETERS GETS 2.5x SPEEDUP COMPARED TO *NEXT*

Method	Params (million)	Times (s)
S-LSTM [9]	**0.26**	0.04
Next [28]	3.95	0.05
LiteNext	3.04	0.03
ST-Attention (ours)	1.98	**0.02**

4) Qualitative Results: We illustrate the qualitative results of *Next* model and our ST-Attention with visualization to make a comparison in Fig. 5. The results demonstrate the effectiveness of our model. When a person meets a couple as shown in Fig. 5(a1) and Fig. 5(a3), ST-Attention is able to pass through the cracks while *Next* model might collide with one of them. In the second row of Fig. 5, we present some scenes where people walk in a group, and ST-Attention is able to jointly predict their trajectories with lower error than *Next* model. We would like to note that in Fig. 5(c2), the predicted path by *Next* model through the wall even though the *Next*

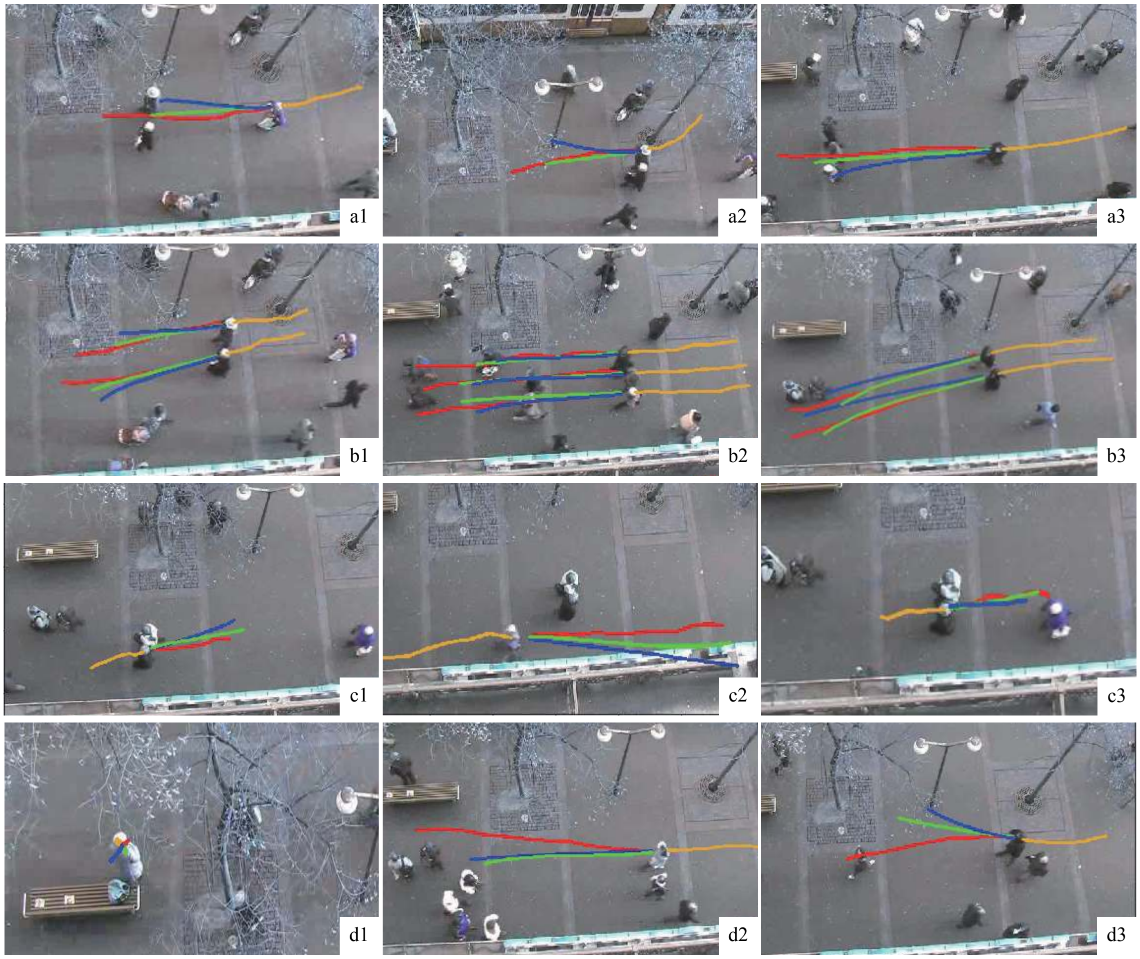

Fig. 5. The visualization results of our ST-Attention predicting path on ETH & UCY datasets: history trajectory (orange), ground truth (red), predicted trajectory for *Next* model (blue) and our model (green). The first three rows show some successful cases and the last row presents some failure examples. we can see that in most cases our predicted trajectory coincides with the ground truth.

model encodes scene semantic information, which testifies that focal attention [29] cannot fully utilize the rich visual feature. In the last row of Fig. 5, several failure cases are shown. In Fig. 5(d1) when a pedestrian waits at the station, he moves slightly as he paces back and forth. ST-Attention assumes he will have a small movement along the previous trend while ground truth has a sudden turn. In Fig. 5(d2) people move in the opposite direction, ST-Attention predicts that target human slows down to avoid collision but actually he gives way toward the right. In Fig. 5(d3) ST-Attention predicts a change of direction toward a wide space whereas the pedestrian goes ahead. Although the predicted paths do not correspond to the ground truth in failure cases, the outputs still belong to acceptable trajectories that pedestrians may take.

D. Experiments on ActEV/VIRAT

ActEV/VIRAT [14] is a benchmark dataset devoted to human activity research. This dataset is natural, realistic and challenging in the field of video surveillance in terms of its resolution and diversity in scenes, including more than 400 videos at 30 frames/s.

1) Implementation Details: In the experiment, training set includes 57 videos and validation set includes 5 videos. Besides, 55 videos are used for testing. In order to keep consistent with the baseline models based on ETH & UCY, activity label is not used in this experiment. Other parameter settings are the same as those in ETH & UCY.

2) Results: Table IV shows quantitative results compared to baseline methods. Combined with Table III, we can see that our model outperforms other methods with lightweight parameters. We also do visualization to reflect the performance of the algorithm intuitively. As show in Fig. 6, our prediction trajectory can better match the ground truth. When a person walks at a normal pace, our model is able to predict his future path, including the situation that the person turns to change the direction of the trajectory such as Fig. 6(b)

TABLE IV
QUANTITATIVE RESULTS ON ACTEV/VIRAT. WE REPORT ADE AND FDE IN PIXEL VALUES

Method	ADE	FDE
Linear	32.19	60.92
S-LSTM [9]	23.98	44.97
Next [28]	19.48	41.45
LiteNext	20.67	44.53
ST-Attention (ours)	**18.39**	**38.11**

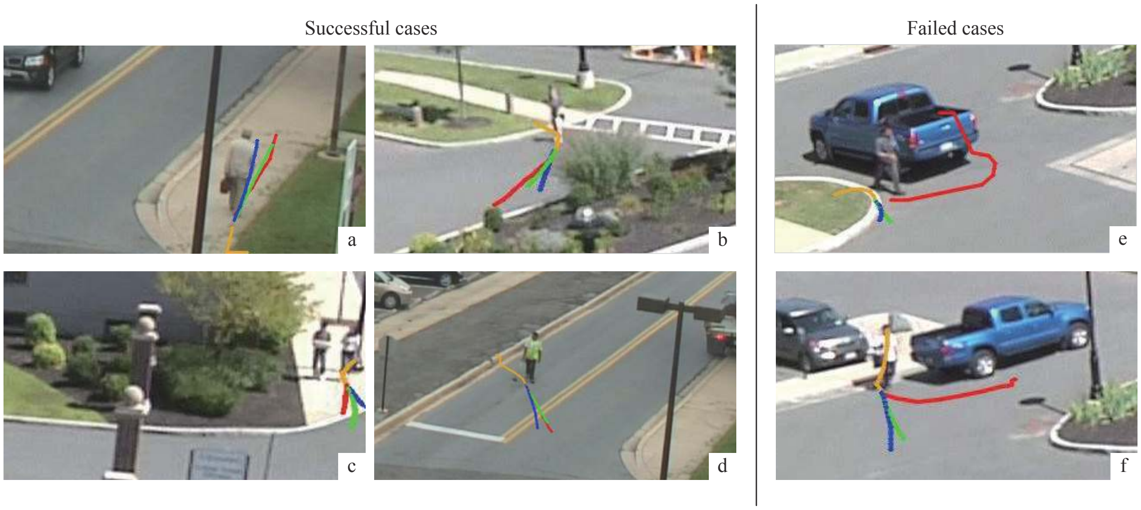

Fig. 6. The visualization results on ActEV/VIRAT dataset: history trajectory (orange), ground truth (red), predicted trajectory for *Next* model (blue) and our model (green).

TABLE V
ABLATION EXPERIMENTS OF THE INTERACTION MODULE AND THE ATTENTION MODULE. WE REPORT ADE AND FDE (ADE/FDE) ON ETH & UCY (METER VALUES) AND ACTEV/VIRAT (PIXEL VALUES)

Datasets	Interaction module		Attention module	
	Without	With	Focal attention	Ours
ETH	0.90/2.02	**0.85/1.85**	0.89/2.00	**0.85/1.85**
HOTEL	0.34/0.71	**0.32/0.66**	0.34/0.69	**0.32/0.66**
UNIV	0.66/1.35	**0.63/1.33**	0.34/0.69	**0.67/1.39**
ZARA1	0.44/0.93	**0.42/0.91**	0.44/0.94	**0.42/0.91**
ZARA2	0.35/0.75	**0.34/0.73**	0.36/0.77	**0.34/0.73**
ActEV/VIRAT	19.30/40.19	**18.39/38.11**	19.88/41.97	**18.39/38.11**

and Fig. 6(c). In Fig. 6(c), the historical trajectory of the target human turns right and then left with a curvature. *Next* model predicts that the human will continue to turn left, but in fact he turns right to the main road as our model predicts. However, our model performs poor when human has a great change of direction due to obvious external interference or other purposes, such as failed cases in Fig. 6. Even though such cases are hard to our model, better performance is achieved compared with other methods.

E. Ablation Study

To explore the role of each module in the trajectory prediction, we make an ablation study on ETH & UCY and ActEV/VIRAT datasets.

1) Effectiveness of the Interaction Module: To verify the importance of interaction module, we train a network removing the interaction branch. Then a comparative experiment with and without interaction module is done and the results are shown in Table V. We can see that a better performance is achieved by the model with interaction module. This is because the interaction module measures the influence of neighbors on the target.

2) Effectiveness of the Attention Module: To evaluate the effectiveness of our attention module, we make a comparison with focal attention [29] which is not learnable. The comparison result is shown in Table V and our attention module performs better than focal attention. This is because our soft attention can automatically learn the weights while focal attention fails, which suggests our attention module is effective for the trajectory prediction.

V. CONCLUSION

In this paper, we present ST-Attention, a spatial-temporal attention model for trajectory prediction. To explore spatial affinity, we use an MLP and the inner product to assign different weights for all pedestrians. To make full use of temporal affinity, the key component named attention model is introduced, which is quite efficient. Our model is fully-differentiable and accurately predicts the future trajectory of the target, which automatically learns the importance for historical trajectories at different time instants and weights the influence of nearby people on the target. Comprehensive experiments on two publicly available datasets have been conducted to demonstrate that ST-Attention can achieve a competitive performance.

Our approach is designed for human trajectory prediction. Future work can extend the model to vehicle trajectory prediction in view of the many similarities between the two predictions mentioned above. Meanwhile we should also distinguish their differences. For example, pedestrians can turn back easily while it is difficult for vehicles, and vehicles can change speed rapidly while pedestrians fail. In particular, it is critical in autonomous driving field to predict the human trajectory jointly with the vehicle trajectory. Besides, intelligent optimization algorithms [41], [42] can be used to learn all the parameters.

References

[1] L. Lv, D. B. Zhao, and Q. Q. Deng, "A semi-supervised predictive sparse decomposition based on task-driven dictionary learning," *Cognitive Computation*, vol. 9, no. 1, pp. 1–10, 2017.

[2] D. B. Zhao, Z. H. Hu, Z. P. Xia, C. Alippi, Y. H. Zhu, and D. Wang, "Fullrange adaptive cruise control based on supervised adaptive dynamic programming," *Neurocomputing*, vol. 125, pp. 57–67, 2014.

[3] D. Li, D. B. Zhao, Q. C. Zhang, and Y. R. Chen, "Reinforcement learning and deep learning based lateral control for autonomous driving," *IEEE Computational Intelligence Magazine*, vol. 14, no. 2, pp. 83–98, 2019.

[4] D. Li, Q. C. Zhang, D. B. Zhao, Y. Z. Zhuang, B. Wang, W. Liu, R. Tutunov, and J. Wang, "Graph attention memory for visual navigation," *arXiv preprint arXiv*: 1905.13315, 2019.

[5] T. Yagi, K. Mangalam, R. Yonetani, and Y. Sato, "Future person localization in first-person videos," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 7593–7602.

[6] D. Makris and T. Ellis, "Path detection in video surveillance," *Image and Vision Computing*, vol. 20, no. 12, pp. 895–903, 2002.

[7] Y. R. Chen, D. B. Zhao, L. Lv, and Q. C. Zhang, "Multi-task learning for dangerous object detection in autonomous driving," *Information Sciences*, vol. 432, pp. 559–571, 2018.

[8] D. B. Zhao, Y. R. Chen, and L. Lv, "Deep reinforcement learning with visual attention for vehicle classification," *IEEE Trans. Cognitive and Developmental Systems*, vol. 9, no. 4, pp. 356–367, 2017.

[9] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, F. F. Li, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 961–971.

[10] A. Gupta, J. Johnson, F. F. Li, S. Savarese, and A. Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 2255–2264.

[11] J. Chen, J. Liu, J. W. Liang, T. Y. Hu, W. Ke, W. Barrios, D. Huang, and A. G. Hauptmann, "Minding the gaps in a video action analysis pipeline," in *Proc. IEEE Winter Applications of Computer Vision Workshops*. IEEE, 2019, pp. 41–46.

[12] S. Pellegrini, A. Ess, and L. Van Gool, "Improving data association by joint modeling of pedestrian trajectories and groupings," in *Proc. European Conf. Computer Vision*. Springer, 2010, pp. 452–465.

[13] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," *Computer Graphics Forum*, vol. 36, no. 3, pp. 655–664, 2007.

[14] G. Awad, A. Butt, K. Curtis, J. Fiscus, A. Godil, A. F. Smeaton, Y. Graham, W. Kraaij, G. Qunot, J. Magalhaes, D. Semedo, and S. Blasi, "Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search," in *TRECVID*. 2018.

[15] G. G. Qu and D. Shen, "Stochastic iterative learning control with faded signals," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 5, pp. 1196–1208, 2019.

[16] Y. R. Chen, D. B. Zhao, and H. R. Li, "Deep Kalman filter with optical flow for multiple object tracking," in *IEEE Int. Conf. Systems, Man, and Cybernetics*. Bari, Italy: IEEE, Oct. 2019. pp. 3036–3041.

[17] C. K. I. Williams, "Prediction with Gaussian processes: From linear regression to linear prediction and beyond," in *Nato Advanced Study Institute on Learning in Graphical Models*. Springer, 1998, pp. 599–621.

[18] D. Helbing and P. Molnár, "Social force model for pedestrian dynamics," *Physical Review E*, vol. 51, no. 5, pp. 4282–4286, 1995.

[19] A. Johansson, D. Helbing, and P. K. Shukla, "Specification of the social force pedestrian model by evolutionary adjustment to video tracking data," *Advances in Complex Systems*, vol. 10, no. supp02, pp. 271–288, 2007.

[20] H. Su, Y. R. Chen, S. W. Tong, and D. B. Zhao, "Real-time multiple object tracking based on optical flow," in *Proc. 9th Int. Conf. Information Science and Technology*. IEEE, 2019. PP. 350–356.

[21] S. Yi, H. S. Li, and X. G. Wang, "Pedestrian behavior understanding and prediction with deep neural networks," in *Proc. European Conf. Computer Vision*. Springer, 2016, pp. 263–279.

[22] S. Y. Huang, X. Li, Z. F. Zhang, Z. Z. He, F. Wu, W. Liu, J. H. Tang, and Y. T. Zhuang, "Deep learning driven visual path prediction from a single image," *IEEE Trans. Image Processing*, vol. 25, no. 12, pp. 5892–5904, 2016.

[23] E. Principi, D. Rossetti, S. Squartini, and F. Piazza, "Unsupervised electric motor fault detection by using deep autoencoders," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 2, pp. 441–451, 2019.

[24] Y. H. Wu, M. Schuster, Z. F. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, and *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," arXiv preprint arXiv: 1609.08144, 2016.

[25] D. Yu and J. Y. Li, "Recent progresses in deep learning based acoustic models," *IEEE/CAA J. Autom. Sinica*, vol. 4, no. 3, pp. 396–409, 2017.

[26] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Machine Learning*. 2015, pp. 2048–2057.

[27] D. Quang and X. H. Xie, "DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences," *Nucleic Acids Research*, vol. 44, no. 11, pp. e107-1–e107-6, 2016.

[28] J. W. Liang, L. Jiang, J. C. Niebles, A. G. Hauptmann, and F. F. Li, "Peeking into the future: Predicting future person activities and locations in videos," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 5725–5734.

[29] J. W. Liang, L. Jiang, L. L. Cao, L. J. Li, and A. Hauptmann, "Focal visual-text attention for visual question answering," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 6135–6143.

[30] Y. Y. Xu, Z. X. Piao, and S. H. Gao, "Encoding crowd interaction with deep neural network for pedestrian trajectory prediction," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 5275–5284.

[31] H. Su, Y. P. Dong, J. Zhu, H. B. Ling, and B. Zhang, "Crowd scene

understanding with coherent recurrent neural networks," in *Proc. 25th Int. Joint Conf. Artificial Intelligence*, vol. 1, pp. 3469–3476, 2016.

[32] H. Su, J. Zhu, Y. P. Dong, and B. Zhang, "Forecast the plausible paths in crowd scenes," in *Proc. 26th Int. Joint Conf. Artificial Intelligence*, vol. 1, pp. 2772–2778, 2017.

[33] P. Zhang, W. L. Ouyang, P. F. Zhang, J. R. Xue, and N. N. Zheng, "SR-LSTM: State refinement for LSTM towards pedestrian trajectory prediction," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 12085–12094.

[34] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, and S. Savarese, "Sophie: An attentive gan for predicting paths compliant to social and physical constraints," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 1349–1358.

[35] C. Wang, H. Han, X. Shang, and X. Zhao, "A new deep learning method based on unsupervised domain adaptation and re-ranking in person re-identification," *Int. J. Pattern Recognition and Artificial Intelligence*, 2019.

[36] M. Köestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*. IEEE, Jun. 2012.

[37] H. Han, M. C. Zhou, and Y. Zhang, "Can virtual samples solve small sample size problem of KISSME in pedestrian re-identification of smart transportation," *IEEE Trans. Intelligent Transportation Systems*, 2019.

[38] H. Han, M. C. Zhou, X. W. Shang, W. Cao, and A. Abusorrah, "KISS+ for rapid and accurate pedestrian re-identification," *IEEE Trans. Intelligent Transportation Systems*, 2020.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*. 2017, pp. 5998–6008.

[40] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *Computer Science*, arXiv preprint arXiv: 1409.0473, 2014.

[41] S. C. Gao, M. C. Zhou, Y. R. Wang, J. J. Cheng, Y. Hanaki, and J. H. Wang, "Dendritic neuron model with effective learning algorithms for classification, approximation and prediction," *IEEE Trans. Neural Networks and Learning Systems*, vol. 30, no. 2, pp. 601–614, 2019.

[42] J. J. Wang, and T. Kumbasar, "Parameter optimization of interval Type-2 fuzzy neural networks based on PSO and BBBC methods," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 1, pp. 247–257, 2019.

Xiaodong Zhao received the B.Eng. degree in automation from North China University of Technology, China in 2018. He is currently working toward the M.A.Sc degree in control science and engineering at University of Science and Technology Beijing. His research interests include trajectory prediction, semantic segmentation, neural architecture search, and autonomous driving.

Yaran Chen received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences in 2018. She is currently an Assistant Researcher at the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China. Her research interests include deep learning, neural architecture search, deep reinforcement learning and autonomous driving.

Jin Guo received the B.S. degree in mathematics from Shandong University, China, in 2008, and Ph.D. degree in system modeling and control theory from the Academy of Mathematics and Systems Science, Chinese Academy of Sciences in 2013. He is currently a Professor with the School of Automation and Electrical Engineering, University of Science and Technology Beijing. His research interests include identification and control of set-valued systems and cyber-physical systems.

Dongbin Zhao (M'06–SM'10–F'20) received the B.S., M.S., Ph.D. degrees from Harbin Institute of Technology, Harbin, China, in 1994, 1996, and 2000, respectively. He was a Postdoctoral Fellow at Tsinghua University, Beijing, China, from 2000 to 2002. He has been a Professor at the Institute of Automation, Chinese Academy of Sciences since 2002, and also a Professor with the University of Chinese Academy of Sciences, China. From 2007 to 2008, he was also a visiting scholar at the University of Arizona. He has published 6 books, and over 90 international journal papers. He serves as the Associate Editor of *IEEE Transactions on Neural Networks and Learning Systems*, *IEEE Transactions on Cybernetics*, *IEEE Transactions on Artificial Intelligence*, *IEEE Computation Intelligence Magazine*, etc. His current research interests include deep reinforcement learning, computational intelligence, autonomous driving, game artificial intelligence, robotics, and smart grids.