Probability Trajectory: One New Movement Description for Trajectory Prediction

Pei Lv[1], Hui Wei[1], Tianxin Gu[1], Yuzhen Zhang[1], Xiaoheng Jiang[1], Bing Zhou[1] and Mingliang Xu[1] (✉)

Abstract Trajectory prediction is a fundamental and challenging task for numerous applications, such as autonomous driving and intelligent robots. Currently, most of existing works treat the pedestrian trajectory as a series of fixed two-dimensional coordinates. However, in real scenarios, the trajectory often exhibits randomness, and has its own probability distribution. Inspired by this observed fact, also considering other movement characteristics of pedestrians, we propose one simple and intuitive movement description, probability trajectory, which maps the coordinate points of the pedestrian trajectory to a two-dimensional Gaussian distribution in space. Based on this unique description, we develop one novel trajectory prediction method, called *Social Probability*. The method combines the probability trajectory and powerful convolution recurrent neural networks together. Both the input and output of our method are probability trajectories, which provide the recurrent neural network with sufficient spatial and random information of moving pedestrians. Furthermore, the *Social Probability* extracts spatio-temporal features directly on the new movement description to generate robust and accurate predicted results. The experiments on public benchmark datasets show the effectiveness of the proposed method.

Keywords Trajectory prediction, Probability trajectory, Convolutional LSTM, Spatio-temporal features.

1 Introduction

The pedestrian trajectory is multimodal, closely related to inherent sense of hearing, vision, touching, thought, personality, and also affected by other factors such as static environments, dynamic human-human interactions, planning destinations, etc. Nevertheless, pedestrians still can intuitively predict the future trajectories of others and adjust themselves in advance. For example, when people walk in shopping malls, streets and stations, they predict the trajectory of others in a short period of time so as to choose their own route at the next moment and avoid collisions. Nowadays, the purpose of trajectory prediction is to enable machines, such as robots, self-driving cars, intelligent tracking systems, to have the ability to predict future trajectories based on historical trajectories. This is one fundamental but extremely challenging task.

In previous works, researchers mainly focused on the following problems in trajectory prediction: interaction among pedestrians [1–6], interaction between pedestrians and scenes [7–9], multi-modality [10, 11]. Especially in recent years, more and more efforts have been made to predict multi-future trajectories [12–16] due to the polymorphism of predicted trajectories. In the real world, the trajectory appears as a probability distribution. When the historical trajectory is known and invariant, a person may have many different future trajectories according to dynamic influencing factors. For example, supposing one person walks twice from the same starting location to the same destination, these two trajectories are usually not exactly the same. Although existing works mentioned above were able to

1 Pei Lv, Hui Wei, Tianxin Gu, Yuzhen Zhang, Xiaoheng Jiang, Bing Zhou and Mingliang Xu are with the School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China (e-mail: ielvpei@zzu.edu.cn; weihui@gs.zzu.edu.cn; txgu@gs.zzu.edu.cn; zyzzhang@gs.zzu.edu.cn; iexhjiang@zzu.edu.cn; iebzhou@zzu.edu.cn; iexumingliang@zzu.edu.cn).

Fig. 1 As illustrated in this scene, two pedestrians are approaching each other. The current position of pedestrians is no longer represented by fixed 2D coordinate points, but a Gaussian probability distribution named after probability trajectory.

predict multi-future trajectories, their inputs were still unimodal, which took the trajectory as two-dimensional coordinate points. Because the movement description is one-dimensional vectors like (x_t, y_t). Obviously, each coordinate point is invariant and nonrandom. These separate points lose the randomness information of the trajectory. Consequently, these methods cannot fully demonstrate the multi-modality of the trajectory caused by inherent randomness.

In addition, earlier works [1–3, 17, 18] have made great progress in modeling the impact of human-human interactions. However, there still exists great challenges since most of these works achieve the purpose of modeling pedestrian interactions by combining hidden states. Because the input is a one-dimensional vector, these hidden states are also one-dimensional, which carries little spatial information. The lack of spatial information makes the problem of modeling interactions complicated and incomprehensible.

In order to solve the above problems, we propose the concept of probability trajectory, which is an intuitive and effective motion description. The way of denoting the trajectory is no longer a series of fixed coordinates, but a probability distribution (Fig. 1). Specifically, we use the probability density function to map the pedestrian coordinate points (x_t, y_t) to two-dimensional Gaussian distributions $G(x_t, y_t)$. Compared with fixed coordinate points, the new movement description avoids the loss of randomness. Moreover, we can conveniently map all pedestrians trajectory at time t into one same two-dimensional space. This gives us a unique advantage in modeling human-human interactions.

Based on the proposed probability trajectory, we further develop one new method called *Social Probability* for predicting robust and accurate pedestrian trajectories. First, the input of our method are probability trajectories, which enables our forecasting model fully consider the randomness of the trajectory. Second, by adding convolution layer to recurrent neural network, our forecasting model can learn spatiotemporal features efficiently. We extract the location information of pedestrians from the two-dimensional probability space through the convolutional neural network. Meanwhile, the two-dimensional probability space contains the location of all pedestrians. Through performing convolution operations on the space, we can extract all pedestrians' location information and easily capture changes in relative location. These two factors are indispensable for modeling interaction.

The rest of the paper is organized as follows. In Section 2, we first review the related work on trajectory prediction. Then we introduce in detail *Social Probability* in Section 3. In Section 4, we further present our experimental results as well as the analysis. Finally, we conclude the proposed method and discuss some future directions in Section 5. To summarize, the main contributions of this paper are as follows:

1. The concept of "probability trajectory" is proposed to denote pedestrian trajectories, which fully demonstrates the inherent randomness of the trajectory to facilitate the subsequent modeling of its multimodal characteristic.

2. The *Social Probability* method is further proposed based on the probability trajectory and recurrent neural networks. With the benefit of convolution operation on the probability trajectories, our method can better model human-human interactions in the visual aspect.

3. The proposed method is successfully tested on public pedestrian datasets. Experiments show that our approach has achieved competitive results both on ADE and FDE compared with other state-of-the-art methods.

2 Related work

2.1 Multimodal trajectory forecasting

In recent years, some researches have tried to model the randomness of trajectory prediction. Gupta et al. [10] solved the trajectory prediction problem using Generative Adversarial Networks (GANs) [19]

and considered the fact that pedestrian trajectories may have multiple plausible predictions. SoPhie [8] combined the scene semantic segmentation model with GANs to model trajectories. Multiverse [11] was a jointly model to generate multiple plausible future trajectories, which contained multi-scale location encodings and convolutional RNNs over graphs. Simultaneously, [14, 15, 20] also proposed probability networks to solve the problem of the randomness in vehicle trajectory prediction. However, these works all treat position information as two-dimensional coordinate points, and input them into the prediction model, which cannot completely describe the random behavior of pedestrians. Different from these works, we take the probability trajectory transformed from the moving trajectory as input, and generate multimodal future trajectories.

2.2 Human-human interaction modeling in trajectory forecasting

For social interaction, researchers utilized multiple methods to establish the interaction model between pedestrians, such as Social Force [2], Social Pooling [1], Attention [3], etc. The methods [21, 22] based on Social Force fully used the principle that attractive forces are used to guide people toward their destinations, and repulsive forces are used to avoid collisions among human-human and human-obstacle. Most of the Social Force-based models try to learn the parameters of the social force functions from real-world crowd datasets. However, Alahi et al. [1] showed the attraction and repulsion alone cannot simulate complex crowd interactions. The approaches [1, 10, 23, 24] used social pooling layer to allow the LSTMs to share their hidden states. This novel design can model human interaction efficiently, but the complexity will increase when the crowd is dense. Then, the methods based on the attention mechanism have emerged [3, 8]. Pedestrains can automatically perceive the importance of certain targets that affects the location at the next time steps. Besides, some methods [25–27] simultaneously learn spatial and temporal interactive patterns to capture spatio-temporal correlation efficiently and comprehensively. These attempts make the interactive model more suitable for real scenarios. RSBG [18] established a group-based social interaction model to explore relationships that are not affected by spatial distance, and Graph Convolutional Neural Network [28] is applied to trajectory prediction. In this paper, the probability trajectory is introduced, and the influence of spatial interaction is automatically
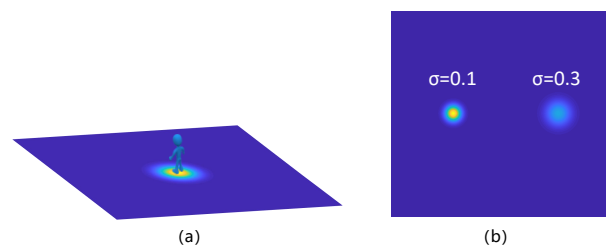

Fig. 2 Probability trajectory describes the location of pedestrians as a probability distribution. (a) shows the distribution of certain trajectory. (b) shows the influential area covered by different σ of the probability distribution.

perceived through convolution operation, which avoids the design of complex interaction modules. The experimental results show that this method has better interaction performance.

2.3 Sequence prediction model

Sequence prediction is to use the sequences from the past to predict the sequences in the future, which is one kind of time series data modeling problem. Convolutional neural network is very useful in the field of computer vision, but is difficult to learn the characteristics of time series data. The recurrent neural network is specially suitable for dealing with the sequence-related data such as audio, video, text. Recurrent neural network and its derivation LSTM [29] and GRU [30] have proved their effectiveness in many fields, such as machine translation [31], text generation [32, 33], speech recognition [34–36], traffic flow prediction [37]. Some researchers have combined convolutional neural network with recurrent neural network, and created novel applications, such as image captioning [33, 38, 39], video understanding [40, 41]. In order to learn spatiotemporal features simultaneously, Shi et al. [42] have added convolution layer to the recurrent neural network. The model called ConvLSTM not only learns the temporal relationship, but also extracts spatial features by convolution layer. We take the advantages of ConvLSTM to obtain spatiotemporal features and directly model the interaction among pedestrians.

3 Our approach

In this section, we first present the new movement description: probability trajectory, which solves the problem of modeling multimodal trajectories from the data description level, then we propose a prediction model based on probability trajectory to describe
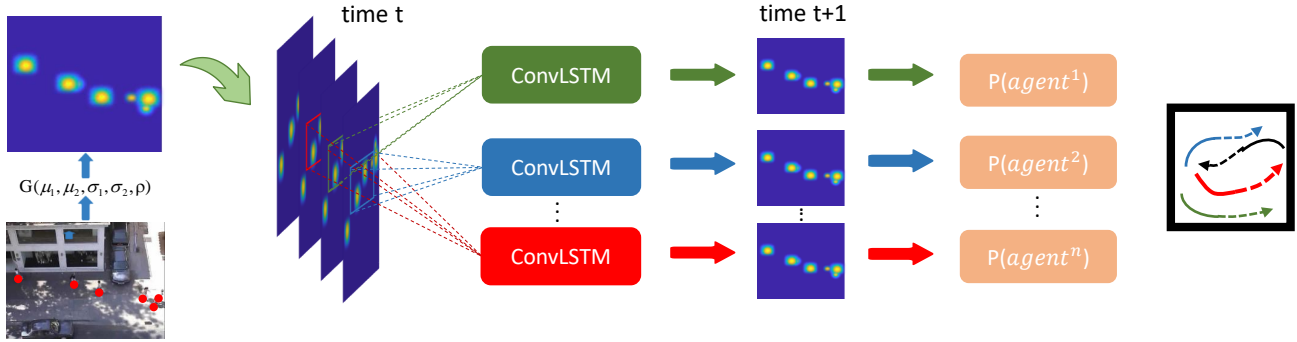
Fig. 3 Overview of our *Social Probability* method. We use a separate ConvLSTM network for each trajectory in the scence. The input and output of the model are all probability trajectories. The probability trajectory is mapped from the fixed coordinate points and one probability trajectory at time t contains the trajectory information of each pedestrian. The ConvLSTM consists of a convolutional layer and gate modules, and has the ability to learn spatiotemporal features. In the predicting stage, the output of our model is also probability trajectories, and trajectory coordinates can be obtained by sampling from them.

human-human interactions conveniently.

3.1   Problem Definition

Our goal is to predict the future trajectories of the pedestrians. The input is the historical location information of each pedestrian in the scene and the output is the trajectory information of all people in the future. Define the historical probability trajectory of the pedestrian as $\mathbf{X} = X_1, X_2, ..., X_n$. The predicted future probability trajectory is denoted as $\widehat{\mathbf{Y}} = \widehat{Y_1}, \widehat{Y_2}, ..., \widehat{Y_n}$, where n represents the number of pedestrians. The input trajectory of a pedestrian i is defined as $X_i \sim N(x_t^i, y_t^i)$ from time steps $t = 1, 2, ..., t_{obs}$ and the future trajectory can be defined similarly as $Y_i \sim N(x_t^i, y_t^i)$ from time steps $t = t_{obs+1}, t_{obs+2}, ..., t_{pred}$, where N represents Gaussian distribution. The prediction is denoted as $\widehat{Y_i}$ and the ground truth is denoted as Y_i.

3.2   Probability Trajectory

3.2.1   Mathematical definition

Supposing the feasible area of the pedestrians is Ω, we represent the location of one pedestrian at time t as a probability distribution on Ω. We use the two-dimensional Gaussian distribution, which can well characterize the location of trajectories. The location distribution at time t has the highest probability density at the center position (x_t, y_t). It means that the location does not have to be at this fixed coordinate points and also has a probability of being located in other area. The farther away from the central location point, the smaller the probability density becomes. Note that, (x_t, y_t) obeys the two-dimensional Gaussian

distribution with parameters $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$:

$$f(x_t, y_t) = (2\pi\sigma_1\sigma_2\sqrt{1-\rho^2})^{-1}\exp[-\frac{1}{2(1-\rho^2)}$$
$$(\frac{(x_t-\mu_1)^2}{\sigma_1{}^2} - \frac{2\rho(x_t-\mu_1)(y_t-\mu_2)}{\sigma_1\sigma_2}) + \frac{(y_t-\mu_2)^2}{\sigma_2{}^2}]$$

$$(1)$$

where μ_1 and μ_2 are the mean value of (x_t, y_t) respectively, σ_1 and σ_2 are the variance of (x_t, y_t), ρ is the correlation coefficient of x_t and y_t. μ_1 is set to x_t and μ_2 is set to y_t. σ_1 and σ_2 are set to 0.3 according to the experience, and ρ is 0.

Using this data structure to represent trajectories, we successfully retain the randomness of trajectories. In two-dimensional space, a pedestrian trajectory is no longer a single point at time t, but a probability distribution. As shown in Fig. 2(a).

3.2.2   Integrate neighbor's information.

At time t, we denote the probability trajectory of pedestrian i as $pMap_t^i$. However, the scene at time t contains multiple pedestrians. Neighboring pedestrians have great influence on the movement decision of each subject pedestrian. In order to enable the model to predict future trajectories based on the location information of the surrounding pedestrians, we need to integrate the probability trajectories of all pedestrians at time t into a same two-dimensional probability space. The probability trajectory at time t denotes $pMap_t$. In two-dimensional space, we integrate $pMap_t^i$ into $pMap_t$ by $\max(\cdot)$ function. Specifically, for the corresponding position in the probability trajectory, we take the larger value as the consolidated value. The formula is as follows.

$$pMap_t = max(pMap_t, pMap_t^i), i = 1, 2, ..., n \quad (2)$$

where n is the number of pedestrians at time t. In order to distinguish the current predicted pedestrian from the rest of the surrounding pedestrians, we set the σ values to 0.1 and 0.3 respectively. The comparison of different σ is shown in Fig. 2(b).

3.3 Convolutional LSTM

Due to its unique structure, the long and short-term memory network (LSTM) has great advantages in processing time sequence data. Moreover, Shi et al. [42] proposed variant of LSTM, which added the convolutional layer to the LSTM module, called ConvLSTM, and proved that the model can learn spatio-temporal information through experiments. Specifically, the main operations are as follows:

$$i_t = \sigma(W_{xi}X_t + W_{hi}h_{t-1} + W_{ci}{}^{\circ}c_{t-1} + b_i) \tag{3}$$

$$f_t = \sigma(W_{xf}X_t + W_{hf}h_{t-1} + W_{cf}{}^{\circ}c_{t-1} + b_f) \tag{4}$$

$$c_t = f_t{}^{\circ}c_{t-1} + i_t{}^{\circ}\tanh(W_{xc}X_t + W_{hc}h_{t-1} + b_c) \tag{5}$$

$$o_t = \sigma(W_{xo}X_t + W_{ho}h_{t-1} + W_{co}{}^{\circ}c_t + b_o) \tag{6}$$

$$h_t = o_t{}^{\circ}\tanh(c_t) \tag{7}$$

where X_t is the input of time t, h_t and c_t are hidden state and cell state, respectively. i_t, f_t, o_t are the gates of the ConvLSTM. They are all 3-dimensional tensors whose last two dimensions are spatial dimensions (width, height). W is the weight matrix. '○' denotes the Hadamard product. At time t, X_t can be input to the module for calculation only when the input gate is activated. Similarly, the past cell state c_{t-1} will be forgotten when the forget gate f_t is activated and the current cell state c_t will be transfered when the output gate o_t is on.

The ConvLSTM uses the current input and past states to determine the future states, and the current input includes not only temporal features, but also spatial features. The temporal features can be learned through the gate structure mentioned above, and the spatial features can be extracted through the convolutional layer embedded in the module. Essentially, trajectory prediction can be regarded as a spatiotemporal sequence generation problem. Therefore, applying ConvLSTM to solve it, we can model the temporal characteristics of the trajectory while also considering the spatial interaction of different trajectories.

3.4 Social Probability

As illustrated in Fig. 3, the *Social Probability* is one trajectory prediction method based on probability trajectories. Firstly, we map the position information of all pedestrians at time t into probability trajectories.

Then, the ConvLSTM module takes two-dimensional probability trajectories as input and outputs predictive probability trajectories at future time $t + 1$. The coordinate points of trajectories can be obtained by sampling from outputs.

3.4.1 Probability-based prediction

The input to the ConvLSTM needs to be two-dimensional tensors. As discussed in Section 3.2, our probability trajectory is a probability distribution in two-dimensional space. Therefore, it is suitable to input probability trajectory into the ConvLSTM model. Moreover, probability trajectories are essentially probability density distributions. The value of the probability trajectory indicates the level of probability density. Modeling probability trajectories directly makes our method a probability-based forecasting method. Our method not only predicts the multimodal future trajectory, but the input historical trajectory is also multimodal, which is different from previous methods. The problem of modeling multimodal features is solved from the data level.

3.4.2 Human-human interactions modeling

The input of our model is probability trajectories of all pedestrians at time t, and is integrated into one two-dimensional space, so modeling human-human interactions is direct and expediently. As illustrated in Fig. 4, after probability trajectories are input into the model, the convolutional layer will extract features in the two-dimensional probability trajectory to obtain the hidden state, which is the feature vector in the RNN-based model. Since the convolution kernel slides across the entire two-dimensional space like a sliding window, hidden states contain the location information of each pedestrian. Namely, due to the convolution operation, the model not only considers the density value of the current position, but also the density value of the surrounding positions when predicting the probability density value at the next time. Therefore, our model considers the location information of all pedestrians at time t, which promotes human-human interactions without complex interaction modules.

3.4.3 Loss function for probability trajectory

We empirically choose the loss function to train our model by refering to the previous works [43, 44]. Since our model focuses on the specific probability density value, rather than some high-dimensional features, such as style, graphics, objects, we use $L2$ loss function to encourage our model to generate accurate probability density distributions.

$$\mathcal{L}_{L2}(\hat{Y}_t^i, Y_t^i) = ||\hat{Y}_t^i - Y_t^i||^2 \tag{8}$$

Method	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
Linear	1.33 / 2.94	0.39 / 0.72	0.82 / 1.59	0.62 / 1.21	0.77 / 1.48	0.79 / 1.59
Vanilla-LSTM	1.09 / 2.14	0.86 / 1.91	0.61 / 1.31	0.41 / 0.88	0.52 / 1.11	0.70 / 1.52
Social-LSTM [1]	1.09 / 2.35	0.79 / 1.76	0.67 / 1.40	0.47 / 1.00	0.56 / 1.17	0.72 / 1.54
Social-GAN [10]	0.81 / 1.52	0.72 / 1.61	0.60 / 1.26	0.34 / 0.69	0.42 / 0.84	0.58 / 1.18
Social-GAN-P [10]	0.87 / 1.62	0.67 / 1.37	0.76 / 1.52	0.35 / 0.68	0.42 / 0.84	0.61 / 1.21
SoPhie [8]	**0.70** / 1.43	0.76 / 1.67	**0.54** / 1.24	**0.30 / 0.63**	0.38 / 0.78	0.54 / 1.15
RSBG [18]	0.80 / 1.53	0.33 / 0.64	0.59 / 1.25	0.40 / 0.86	**0.30 / 0.65**	**0.48** / 0.99
NEXT [7]	0.73 / 1.65	**0.30 / 0.59**	0.60 / 1.27	0.38 / 0.81	0.31 / 0.68	**0.46** / 1.00
Ours	0.74 / **1.22**	0.49 / 0.85	0.63 / **1.23**	0.42 / 0.78	0.38 / 0.70	0.53 / **0.95**

Tab. 1 Quantitative results of different methods on ETH, Hotel (from ETH) and UNIV, ZARA1, ZARA2 (from UCY) datasets. We show two error metrics Average Displacement Error (ADE) and Final Displacement Error (FDE) on the task of predicting 12 future time steps.
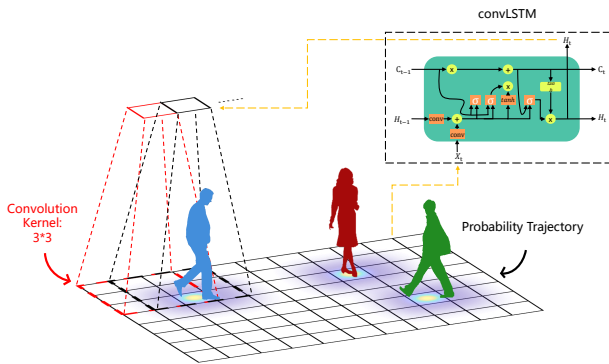

Fig. 4 The convolution layer added into LSTM makes the model have the ability to extract spatial features. Convolution operation on the probability trajectory enables our method to model human-human interactions more intuitively and efficiently.

Here, $\hat{Y}_t^i$ and Y_t^i are predicted and ground truth probability trajectory for person i at time t respectively.

4   Experiment

In this section, we show the experimental results on five public datasets, and compare them with current state-of-the-art methods, then analyze the performance of our method.

4.1   Datasets

We validate the proposed model on the public datasets ETH [22] and UCY [45], which are the widely used benchmark datasets in the field of trajectory prediction. Most of the current state-of-the-art methods are evaluated on these datasets. It contains a total of 1536 labeled pedestrians in 4 different scenes. These datasets are based on binocular vision for the research

of pedestrian trajectory tracking and prediction. There are totally 5 sub datasets, where ETH contains two sub datasets as ETH and HOTEL, and UCY consist of three parts as ZARA1, ZARA2 and UNIV. Similar to the previous works, we still observe the historical trajectory for the past 8 time steps (3.2 seconds) and predict the future trajectory for the next 12 time steps (4.8 seconds).

4.2   Evaluation Metrics and Methods

According to previous works [1], we use two evaluation metrics.

1. *Average displacement error (ADE):* The average Euclidean distance between the predicted trajectories and the true trajectories at each prediction time step.
2. *Final displacement error (FDE):* The Euclidean distance between the predicted destination and the ground truth destination at the last prediction time step.

The two evaluation scales are defined as:

$$ADE = \frac{\sum\limits_{i \in \mathbb{Z}} \sum\limits_{t=T_{obs}+1}^{T_{pre}} \sqrt{((\hat{x}_t^i, \hat{y}_t^i) - (x_t^i, y_t^i))^2}}{\mathbb{Z} * T_{pre}} \quad (9)$$

$$FDE = \frac{\sum\limits_{i \in \mathbb{Z}} \sqrt{((\hat{x}_{T_{pre}}^i, \hat{y}_{T_{pre}}^i) - (x_{T_{pre}}^i, y_{T_{pre}}^i))^2}}{\mathbb{Z}} \quad (10)$$

where $(\hat{x}_t^i, \hat{y}_t^i)$ and (x_t^i, y_t^i) are the predited and ground truth coordinates for pedestrian i at time t, $\mathbb{Z}$ is the total number of pedestrian in the testing set.

We use a leave-one-out approach to evaluate the performance of the model. Four sets are used as the training set and verification set, and the remaining one is used as the testing set to gain the results.

4.3 Implementation Details

The number of layers of the ConvLSTM model is 5 and the channel dimension of the hidden state in each layer is 128, 64, 64, 32, 32 respectively. The kernel size of the convolutional layer is 3*3 and the padding is 1. We train our model using Adam [46] with the initial learning rate of 0.001. The size of the probability trajectory and the hidden state of our model are both 100*100. In the stage of prediction, the variance of the current pedestrian to be predicted is set to 0.1, and the other pedestrians are set to 0.3. In the testing stage, we sample 20 times from the probability trajectory predicted by the model, and select the best prediction in Euclidean distance for quantitative estimation.

4.4 Compared With different Methods

As shown in Table 1, we choose the following methods for comparison:

1. Linear: A linear regression model to predict the trajectory by minimizing the least square error.
2. Vanilla-LSTM: Use the LSTM model to predict the future trajectory. This method only considers its own historical trajectory and does not consider any other factors.
3. Social-LSTM [1]: The social-pooling layer is added to LSTM, so that the model has the ability to model human-human interactions.
4. Social-GAN [10]: A trajectory prediction model trained with GAN architecture is designed to improve existing models in terms of rationality, diversity, and prediction speed. The model pays attention to the feasibility of predictive generation trajectory in social rules.
5. Social-GAN-P [10]: The only difference with the Social-GAN is that the pooling mechanism is not applied.
6. SoPhie [8]: An interpretable framework based on GAN for trajectory prediction. It uses two information sources, the historical trajectory of all pedestrians in a scene and the scene context information of the scene image.
7. RSBG [18]: A group-based social interaction model to explore pedestrian relationships that are not affected by spatial distance. Graph Convolutional Neural Network is applied to trajectory prediction in this model.

4.5 Quantitative Analysis

Table 1 lists two error metrics Average Displacement Error (ADE) and Final Displacement Error (FDE) of our method against existing methods, with the

Method	ADE	FDE
Our method with attention	0.99	1.73
Our method without attention	0.91	1.61

Tab. 2 Ablation study for attention mechanism. We show the influences when adding attention to our methods. The experiment was conducted on ETH dataset without data augmentation.

task of predicting 12 future time steps according to 8 historical time steps. We follow the comparative works to choose the best prediction among multiple samples for quantitative analysis. It can be seen that the linear model usually performs the worst. Because it is only suitable to predict the straight trajectories, and is insensitive to pedestrian interaction. Social-LSTM and Social-GAN perform better than Linear method since they can handle interacrions among pedestrians by introducing the corresponding interactive module. We can see that our method outperfroms all others on FDE of the ETH and UNIV datasets, avoiding more potential collisions in the future. Although the performance of our method is not the best on other datasets, they are still very competitive and significantly superior to the linear model except for the HOTEL dataset with a small amount of pedestrians. Therefore, it fully proves that our method own inherent interaction function without complex interaction modules. In addition, compared with ADE, our method performs better in FDE, especially in the ETH dataset. This reflects that our method has more advantages in the aspect of predicting destinations.

4.6 Qualitative Analysis

Fig. 5 shows the positive and negative samples on each dataset. The blue trajectories are the predictive trajectories of future 12 time steps from the observed trajectories of past 8 time steps marked as green, and the red trajectories are the ground truth trajectories. The scenarios in the ETH, HOTEL, UNIV, ZARA1 and ZARA2 dataset are shown in column 1 to 5. The visualization results show that our model is able to correctly predict the future path and have the ability to model human-human interactions. According to the positive samples, the model can avoid obstacles in advance when interacting with others. Besides, our method is also suitable for crowded scenes, when multiple people are walking forward to the same or different direction, voiding each other by following or interpolation.
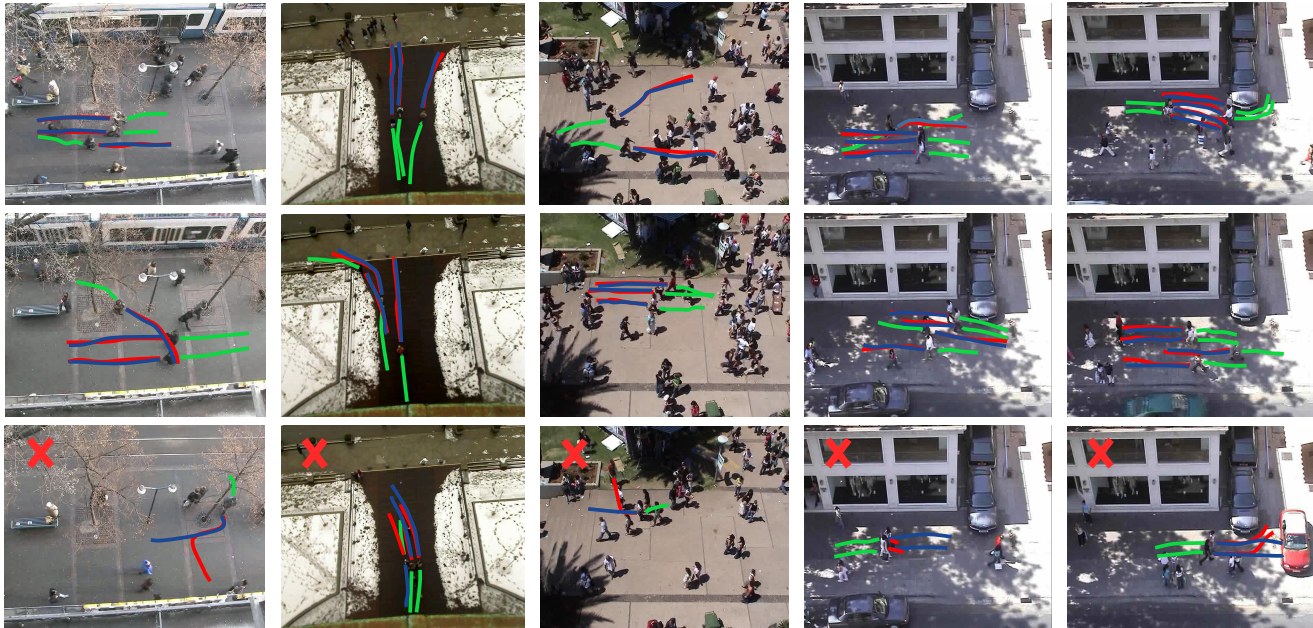
Fig. 5 Visualization of qualitative analysis. Given the observed trajectories of past 8 time steps(Green), our method predicts the trajectories of future 12 time steps(Blue), and the ground truth trajectories(Red) is compared with the predicted trajectory. The results for ETH, HOTEL, UNIV and ZARA1 and ZARA2 are shown in column one to five, respectively. On the first two row, we show examples where our method successfully predicts the trjectories with small errors. The last row shows some failure cases. Note that, these are representative examples selected from the visual results.

The last row shows some negative samples. These examples have large gaps in predictions, or have the wrong direction. By analyzing the source videos, we found that these failure cases were generated when pedestrians stopped walking or turned suddenly. The main reason for these results is the unpredictability of pedestrian intentions. Another reason is that when pedestrians interact with the physical surrounding, the model cannot well perceive the scene information. Our method has not yet covered the integration of scene information which is the direction of our future research.

4.7 Ablation Study

4.7.1 Attention mechanism

During the experiment, we try to use spatial attention mechanism [47] to improve the prediction accuracy of our model. In the two-dimension space of probability trajectory, attention module is applied to capture which location has more influence. However, we found that the attention mechanism did not improve our experimental results as expected. The comparative experiments are shown in Table 2. The reason may be that the probability trajectory has already played a role in attention. The probability density of each spatial position represents the importance of the location,

Method	ADE	FDE
Our method (full algorithm)	0.74	1.22
Our method without integration	0.86	1.64

Tab. 3 Ablation study for our full method without the integration of neighbor's probability trajectories. The experiment was conducted on ETH dataset.

namely the weight value in the attention mechanism.

4.7.2 The integration of probability trajectories

As discussed in Section 3.2, we integrate the probability trajectories of all pedestrians at time t into one same two-dimensional probability space. In this section, we remove the integration to verify the ability to model interaction of our method. When predicting the trajectory of the person i, the probability trajectory only contains its own trajectory information. The trajectory information of people around is not integrated into his probability trajectory. We conduct experiments on the ETH dataset and the results are shown in Table 3. We can find that the method with integration reaches an improvement of 14.0 % and 25.6 % in ADE and FDE. It is fully proved that the integration of the probability trajectories has the ability to model human-human interactions.
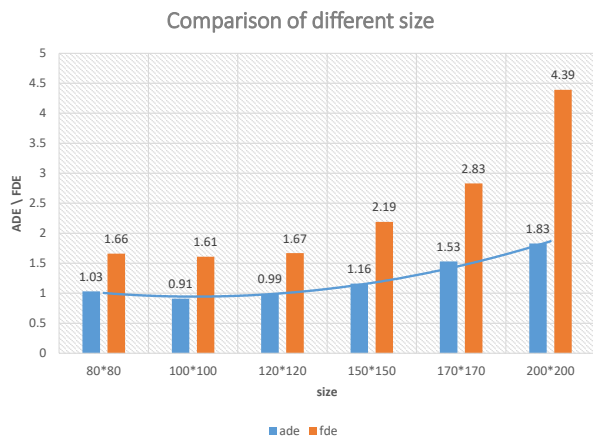
Fig. 6 Ablation study for different sizes of probability trajectories. The blue line shows the trend of ADE. Experiments was conducted on ETH dataset without data augmentation.

4.7.3 The size of probability trajectory

The probability trajectory is two-dimensional, so the suitable size is necessary. We set up the comparison experiment, and the size is 80*80, 100*100, 150*150, 200*200. The results are shown in Fig. 6. From the figure, we can find that when the size is 100*100, the predicted result is the best. Too large or too small will cause decreases in prediction accuracy. We sampled from the ground truth and found that as the size increases, the sampling error also increases. The sampling error may be the reason for the decrease of prediction accuracy. On the contrary, as the size decreases, the model is not capable to modeling large enough amounts of data.

5 Conclusions

In this paper, we propose the concept of probability trajectory, which has more advantages in representing the randomness of trajectories, and explore a new trajectory prediction method based on it. To encode social interaction features, we introduce ConvLSTM, a sequence to sequence prediction model, which has the ability to model spatiotemporal characteristics. Experiments on public datasets show the effectiveness of our method. Although it is not state-of-the-art in all datasets, our method is simple and has great potential. In addition, our current work does not incorporate physical surrounding, but it is obvious that adding physical scenarios to our model is straightforward and convenient, and this is the direction of our future work.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016.

[2] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995.

[3] Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds. In *2018 IEEE international Conference on Robotics and Automation (ICRA)*, pages 1–7. IEEE, 2018.

[4] Shuai Yi, Hongsheng Li, and Xiaogang Wang. Pedestrian behavior understanding and prediction with deep neural networks. In *European Conference on Computer Vision*, pages 263–279. Springer, 2016.

[5] Pu Zhang, Wanli Ouyang, Pengfei Zhang, Jianru Xue, and Nanning Zheng. Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12085–12094, 2019.

[6] Mingliang Xu, Chunxu Li, Pei Lv, Nie Lin, Rui Hou, and Bing Zhou. An efficient method of crowd aggregation computation in public areas. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2814–2825, 2017.

[7] Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander G Hauptmann, and Li Fei-Fei. Peeking into the future: Predicting future person activities and locations in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5725–5734, 2019.

[8] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1349–1358, 2019.

[9] Sai-Keung Wong, Yu-Shuen Wang, Pao-Kun Tang, and Tsung-Yu Tsai. Optimized evacuation route based on crowd simulation. *Computational Visual Media*, 3(3):243–261, 2017.

[10] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018.

[11] Junwei Liang, Lu Jiang, Kevin Murphy, Ting Yu, and Alexander Hauptmann. The garden of forking paths: Towards multi-future trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10508–10518, 2020.

[12] Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. *arXiv preprint arXiv:1910.05449*, 2019.

[13] Yuke Li. Which way are you going? imitative decision learning for path forecasting in dynamic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 294–303, 2019.

[14] Osama Makansi, Eddy Ilg, Ozgun Cicek, and Thomas Brox. Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7144–7153, 2019.

[15] Charlie Tang and Russ R Salakhutdinov. Multiple futures prediction. In *Advances in Neural Information Processing Systems*, pages 15424–15434, 2019.

[16] Hao Xue, Du Q Huynh, and Mark Reynolds. Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1186–1194. IEEE, 2018.

[17] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6272–6281, 2019.

[18] Jianhua Sun, Qinhong Jiang, and Cewu Lu. Recursive social behavior graph for trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 660–669, 2020.

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[20] Luca Anthony Thiede and Pratik Prabhanjan Brahma. Analyzing the variety loss in the context of probabilistic trajectory prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9954–9963, 2019.

[21] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 935–942. IEEE, 2009.

[22] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th International Conference on Computer Vision*, pages 261–268. IEEE, 2009.

[23] Hang Su, Jun Zhu, Yinpeng Dong, and Bo Zhang. Forecast the plausible paths in crowd scenes. In *IJCAI*, volume 1, page 2, 2017.

[24] Yanyu Xu, Zhixin Piao, and Shenghua Gao. Encoding crowd interaction with deep neural network for pedestrian trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5275–5284, 2018.

[25] Qicheng Tang, Mengning Yang, and Ying Yang. St-lstm: A deep learning approach combined spatio-temporal features for short-term forecast in rail transit. *Journal of Advanced Transportation*, 2019, 2019.

[26] Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. Gman: A graph multi-attention network for traffic prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1234–1241, 2020.

[27] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017.

[28] Mikael Henaff, Joan Bruna, and Yann LeCun. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*, 2015.

[29] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[30] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[31] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning

to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[32] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897, 2014.

[33] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[34] Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. End-to-end continuous speech recognition using attention-based recurrent nn: First results. *arXiv preprint arXiv:1412.1602*, 2014.

[35] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In *Advances in neural information processing systems*, pages 2980–2988, 2015.

[36] Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772, 2014.

[37] Bailin Yang, Shulin Sun, Jianyuan Li, Xianxuan Lin, and Yan Tian. Traffic flow prediction using lstm with feature enhancement. *Neurocomputing*, 332:320–327, 2019.

[38] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[39] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016.

[40] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[41] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852, 2015.

[42] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810,

2015.

[43] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2235–2245, 2018.

[44] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, 2016.

[45] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, volume 26, pages 655–664. Wiley Online Library, 2007.

[46] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[47] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

Pei Lv received the Ph.D. degree from the State Key Laboratory of CAD&CG, Zhejiang University Hangzhou, China, in 2013. He is an Associate Professor with the School of Information Engineering, Zhengzhou University, Zhengzhou, China. His research interests include computer vision and computer graphics. He has authored more than 30 journal and conference papers in the above areas, including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, CVPR, ACM MM, and IJCAI.
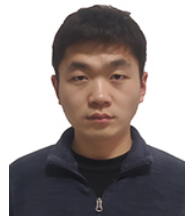
Hui Wei received the B.S. degree from the Software Engineering Department, Henan Agricultural University, China, in 2018. He is currently a master student in the School of Information Engineering of the Zhengzhou University. His research interests include computer vision, trajectory prediction.

Tianxin Gu received the B.S. degree from the Network Engineering Department, Henan Polytechnic University, China, in 2018. She is currently a master student in the School of Information Engineering of the Zhengzhou University. His research interests include computer vision, trajectory prediction.

Yuzhen Zhang received the M.S. degree from the Software Engineering Department, Henan Polytechnic University, China, in 2019. She is currently a doctor student in the school of Information Engineering of the Zheng Zhou University. Her research interests include computer vision and trajectory prediction.

Xiaoheng Jiang received the B.S., M.S., and Ph.D. degrees in electronic information engineering from Tianjin University, Tianjin, China, in 2010, 2013, and 2017, respectively. He is currently a Lecturer with the School of Information Engineering, Zhengzhou University, Zhengzhou, China. His research interests include computer vision and deep learning.

Bing Zhou received the B.S. and M.S. degrees in computer science from Xi'an Jiao Tong University, Xi'an, China, in 1986 and 1989, respectively, and the Ph.D. degree in computer science from Beihang University, Beijing, China, in 2003. He is currently a Professor with the School of Information Engineering, Zhengzhou University, Zhengzhou, China. His research interests include video processing and understanding, surveillance, computer vision, and multimedia applications.
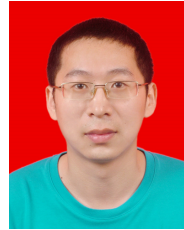
Mingliang Xu received the Ph.D. degree in computer science and technology from the State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou, China, in 2012. He is a Full Professor with the School of Information Engineering, Zhengzhou University, Zhengzhou, China, where he is currently the Director of the Center for Interdisciplinary Information Science Research and the Vice General Secretary of ACM SIGAI China. His research interests include computer graphics, multimedia, and artificial intelligence. He has authored more than 60 journal and conference papers in the above areas, including ACM Transactions on Graphics, ACM Transactions on Intelligent Systems and Technology, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, ACM SIGGRAPH (Asia), ACM MM, and ICCV.