# IA-LSTM: Interaction-Aware LSTM for Pedestrian Trajectory Prediction

Jing Yang, *Member, IEEE,* Yuehai Chen, *Member, IEEE,* Shaoyi Du, *Member, IEEE,* Badong Chen, *Senior Member, IEEE,* and Jose C. Principe, *Fellow, IEEE,*

***Abstract*—Predicting the trajectory of pedestrians in crowd scenarios is indispensable in self-driving or autonomous mobile robot field because estimating the future locations of pedestrians around is beneficial for policy decision to avoid collision. It is a challenging issue because humans have different walking motions and the interactions between humans and objects in the current environment, especially between human themselves, are complex. Previous researches have focused on how to model the human-human interactions, however, neglecting the relative importance of interactions. In order to address this issue, we introduce a novel mechanism based on the correntropy, which not only can measure the relative importance of human-human interactions, but also can build personal space for each pedestrian. We further propose an Interaction Module including this data-driven mechanism that can effectively extract feature representations of dynamic human-human interactions in the scene and calculate corresponding weights to represent the importance of different interactions. To share such social messages among pedestrians, we design an interaction-aware architecture based on the Long Short-Term Memory (LSTM) network for trajectory prediction. We demonstrate the performance of our model on two public datasets and the experimental results demonstrate that our model can achieve better performance than several latest methods with good performance.**

***Index Terms*—pedestrian trajectory prediction, human-human interactions, correntropy, long short-term memory network.**

## I. Introduction

TRAJECTORY prediction of pedestrians is of major importance for applications in various fields including autonomous driving [1], robot navigation [2]–[5] and surveillance camera analysis [6], [7]. Estimating the future positions of pedestrians in real world accurately is necessary and beneficial for some tasks [3], [8]–[14], such as robot-assisted pedestrian regulation [13] and person reidentification [14]. Forecasting pedestrians motions, however, is an extremely difficult problem, due to the complex interactions between pedestrians. For example, as we move in an environment, we would yield right-of-way to nearby people [15], tend to walk in groups [16] and follow social norms, change our trajectory based on changes of other pedestrians [17]. A main challenge of trajectory forecasting lies in how to incorporate human-human interaction into trajectory prediction of pedestrians.

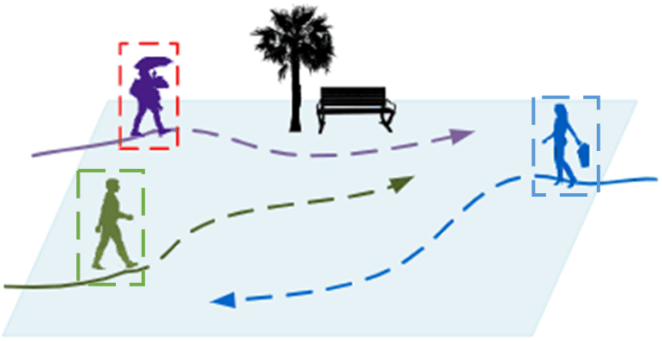


Fig. 1. An illustration of a common scene where static interactions and dynamic interactions both occur. The pedestrian who is under an umbrella (framed by red dashed line) has static interactions with stationary obstacles (tree and bench), dynamic interactions with two other pedestrians, which has influences on his/her future trajectory.

One of the most classic approaches having considered both static and dynamic interactions, is a named "Social Force" model proposed by D. Helbing and P. Molnar [18]. This approach assumes that the objects (including humans) in the scene could exert forces upon the target pedestrian to influence his/her movements. These interactions are cleverly modeled by defining these different repulsive forces. The work [9] viewed pedestrians as decision-making agents which also consider a plethora of personal, social, and environmental factors to decide where to go next. Yang et al. built two datasets for pedestrian motion models that consider both interpersonal and vehicle crowd interaction [10]. With tracking multiple people in complex scenarios, a powerful dynamic model [3] is still competitive in some datasets nowadays [1]. However, these forces are calculated based on hand-crafted functions, a fixed form of physical model simulating changes in behavior, which means this model can only capture simple dynamic interactions and the complex crowd social behaviors are always neglected [19].

Recently, with huge advancements of deep learning, Recurrent Neural Networks (RNN) and its variants such as Long Short-Term Memory (LSTM) [20] and Gated Recurrent Units (GRU) [21] are widely applied in time series problems including pedestrian trajectory prediction [17], [19], [22]–[27]. Most of these works [17], [19], [22], [27] have quantified the

This work was supported by the National Natural Science Foundation of China under Grant Nos. 62073257 and 61971343, and the Natural Science Basic Research Plan in Shaanxi Province of China under Grant No. 2020JM-012. (*Corresponding author: Shaoyi Du.*)

Jing Yang and Yuehai Chen are co-first author.

Jing Yang and Yuehai Chen are with the Institute of Control Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: cyh0518@stu.xjtu.edu.cn; jasmine1976@xjtu.edu.cn).

Shaoyi Du and Badong Chen are with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: dushaoyi@gmail.com; chenbd@xjtu.edu.cn ).

Jose C. Principe is with the Computational NeuroEngineering Laboratory, Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA (e-mail: principe@cnel.ufl.edu).

human-human interactions in the scene and jointly predict the trajectories. The "Social Attention" model [27] has considered the relative importance of the interactions and measure these different interactions by defining a score function based on the *scaled dot product attention mechanism* [28]. However, this mechanism introduced extra parameters to the original model, and the more pedestrians in the scene, the greater the amount of computation, which greatly increased training costs. The "Social LSTM" model [19] assumes that the human-human interactions caused by different pedestrians in the scene are equally important, which did not consider the real situation thoroughly.

The pedestrian trajectory prediction which need to consider human-human interactions for avoiding collisions with other pedestrians, is different to the behavior of other sequences task [29], [30]. In the real world, there are different interactions among different pedestrians. As shown in Figure 1, pedestrian (framed by red dashed line) pays more attention to the interaction caused by nearby person (framed by green dashed line), the interaction caused by the person (framed by blue dashed line) walking on the distant street is definitely less concerned about. However, considering only neighboring pedestrians [19] is also insufficient to capture the dynamic human-human interactions. A plenty of researches reveal that human has his/her own personal space [19], [31]–[34]. Human respect personal space: once a pedestrian enters the personal space of other people, the generated interaction is huge, but it will drop sharply outside this personal space. This type of human-human interactions cannot be measured just by the relative Euclidean distance between the pedestrians, which is the most common measurement method in recent researches [9]. Accurate measurement of human-human interactions is the key issue of pedestrian trajectory prediction.

As discussed above, there are different interactions among different pedestrians, and dynamic human-human interactions have an unique step property which is related to everyone's personal space. In this paper, to address the argued above problems, we propose a novel mechanism based on the correntropy to measure the relative importance of the human-human interactions and represent everyone's personal space. In this mechanism, the value of the correntropy between two pedestrians is calculated and used as the weight of their interactions, the personal space is defined through setting suitable kernel of correntropy. Then, we define an "Interaction Module" to extract feature representations of the human-human interaction and attach the weight to it. Finally, the social information of such human-human interactions are shared through our designed a variant structure of LSTM.

The correntropy is calculated from original data rather than defining hand-crafted function with specific settings, and it is capable of generalizing high dimensional features combining the time structure and statistical distribution, which is suitable in this case. Our designed Interaction-Aware LSTM architecture is capable of capturing different complex human-human interactions among crowds through our proposed "Interaction Module" autonomously without any additional notations or social rules, and then these interactions are shared among pedestrians in the scene for future trajectory prediction. Finally, we use two public datasets ETH [35] and UCY [1] for experiments and experiments demonstrate that our model can achieve better performance than several latest methods with good performance. We also analyze the trained model to understand the human-human interactions in a social way.

The reminder of this paper is as follows. We provide a brief introduction of the LSTM networks and correntropy in Section 2. In Section 3, the proposed model is detailed described. The performance of our model approach is subsequently verified in Section 4. In last Section 5, we analyze the predicted trajectories to demonstrate the behavior of pedestrians learned by our model.

## II. Related Works and Preliminary

### A. Methods for trajectory prediction of pedestrians

**Traditional Methods.** Some previous traditional researches mainly focus on how to model the real human among the crowds and use this descriptive model to predict the future trajectories. For instance, some early works are based on Bayesian model, using online Bayes filters (Kalman filters and particle filters) [36] or using a dynamic Bayesian network (DBN) [37], [38]. These models make predictions without considering whether static or dynamic interactions in the scene and the predicted results are highly deviated from ground truth.

Another kind of traditional approaches has considered the static interactions (static texture) in the scene using a grid graph to predict the entire future positions at the same time. These approaches design a grid graph to represent all possible paths of one pedestrian and assign different weights to the edges, therefore the prediction problem becomes to the shortest path problem. In particular, Huang et al. [39] and Walker rt al. [40] use appearance texture of the scene to build the cost map. Xie et al. [41] uses objects in the scene to define the cost map. Lack of considering dynamic human-human interactions in the scene makes these approaches inappropriate in crowd scenarios.

**Deep Learning Methods.** Recently, with huge advancements of deep learning, numerous methods are applied to address this task and achieve better results. For instance, Convolutional Neural Networks (CNNs) are applied by Yi et al. [42] to make predictions. They proposed a named behavior-CNN which has three dimensional data channels with bias maps to consider different behaviors at specific locations such as the entrances and obstacles occurred in the scene. With the limitation of the input form, this method has only considered the static interactions in the scene while neglecting the influence caused by other pedestrians.

In addition, Inverse Reinforcement Learning (IRL) [43], [44] was first presented to solve the optimal motion planning of robots [45], and Kitani et al. [46] cleverly introduced IRL to trajectory prediction. Instead of directly predicting the future path, they teach the robot take actions sequentially and obtain its entire future positions as the predicted path. Lee et al. [47] also applied this method to predict the moving trajectory of the football player and Wei et al. [48] proposed an approach based on the fictitious play theory to predict trajectories of multiple pedestrians with several predefined goals. In general, these

works mentioned above are named as activity forecasting. This kind of methods converts static and dynamic interactions to what the target robot observed, however, the most difficult part is how to define the reward, namely the effect of these different interactions.

### B. Correntropy

Correntropy, the abbreviation for correlation entropy, proposed by Principe et al. [49], [50], is an effective kernel-based similarity measure in feature space [51] based on the information theoretic learning(ITL) [52] with various applications such as classification [53]–[55], regression [56], [57], deep learning [53], [58], pitch detection in speech [59], adaptive filtering [60], [61], principal component analysis (PCA) [62], [63], and so on.

For two different random variables $X$ and $Y$,the correntropy between $X$ and $Y$ is defined as:

$$V(X,Y) = \mathbf{E}[\kappa(X,Y)] = \int \kappa(x,y) dF_{\mathrm{XY}}(x,y) \qquad (1)$$

where $E$ is the expectation operator, $\kappa(x,y)$ is a shift-invariant Mercer kernel, and $F_{\mathrm{XY}}(x,y)$ denotes the joint distribution function of $(X,Y)$. The most popular kernel used in correntropy is the Gaussian kernel:

$$\kappa_\sigma(x,y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\|x-y\|^2/2\sigma^2\right) \qquad (2)$$

where the $\sigma > 0$ denotes the kernel size (or kernel bandwidth).

Human respect personal space and yield right-of-way. Each pedestrian has their comfortable personal space, once others enter this space, pedestrian would feel uncomfortable. In other words, generated interactions are huge in the range of personal space, and would drop sharply outside this personal space. Moreover, pedestrian is easier effected by nearby human, rather than distant human. In this case, we could consider the distant pedestrians as outliers.

Inspired by the great properties of correntropy, we extend correntropy to the trajectory prediction, namely, to build personal space for each pedestrian and measure the relative importance of the human-human interactions in the scene. More specifically, we first build the personal space of pedestrians through designing suitable Gaussian kernel. Then, we use the robustness to outliers of correntropy, to model human-human interactions. Compared with other similarity measures, such as mean-square error (MSE), the correntropy has the ability to generalize the conventional correlation function to high dimensional feature spaces and it is robuster to impulsive noises or outliers [62], [64]–[67]. Correntropy defines a non-homogeneous metric, which can be used as an outlier-robust error measure in robust signal processing [64]. In similar way, Du et al. proposed a robust graph-based method based on correntropy to deal with labeling noisy data [65]. The correntropy which is calculated directly from original data without additional settings, is naturally suitable in such pedestrians trajectory prediction case.

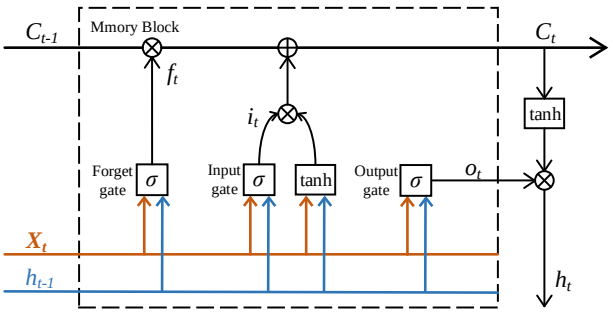



Fig. 2. The architecture of LSTM, which contains forget gate, input gate, and output gate.

### C. Vanilla Long Short-Term Memory Networks

Long Short-Term Memory Networks (LSTM) is a recurrent neural networks variant proposed by Sepp Hochreiter and Jürgen Schmidhuber [20] in order to solve the gradient vanishment or gradient explosion issue in conventional Back-Propagation Through Time (BRBT) [68] and Real-Time Recurrent Learning (RTRL) [69]. Unlike traditional RNNs [70], an LSTM network is well-suited to learn from experience to classify, process and predict time series when there are very long time lags. LSTM has various applications in dealing with the time series problem such as Natural Language Process [71]–[77] including Sentiment Analysis [75] and Air-Quality Prediction [77], Image Generation [78]–[80], Fault Diagnosis [81], Machine Translation [82] and so on.

As shown in Figure 2, the LSTM architecture includes a memory block with cell, input gate, forget gate and output gate. The core component of LSTM is the cell which can store the long range contextual information and these gates are designed to remove or add the information to cell state. Given the input vector $X_t$ at time $t$, cell state $C_{t-1}$ and hidden state $h_{t-1}$ at time $t-1$, the forget gate's activation vector $f_t$, the input gate's activation vector $i_t$ and the output gate's activation vector $o_t$ are firstly computed as follows:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \qquad (3)$$

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \qquad (4)$$

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right) \qquad (5)$$

where $\cdot$ represents matrix multiplication, $W_f$, $W_i$ and $W_o$ are weight matrices of the three gates, respectively, and $b_f$ , $b_i$ and $b_o$ are their bias vectors. $\sigma$ is the gate activation function. Then, we obtain the cell state from Equation 6:

$$C_t = f_t * C_{t-1} + i_t * \tanh\left(W_c\left[h_{t-1}, x_t\right] + b_c\right) \qquad (6)$$

and then hidden state from Equation 7:

$$h_t = o_t * \tanh\left(C_t\right) \qquad (7)$$

where $*$ represents point-wise multiplication. The back feedback progress to update the parameters is the same with RNN and can be found in [20].

Considering that correntropy is robust to impulsive outliers, we designed a novel mechanism based on the correntropy to accurately model the interaction between people in the

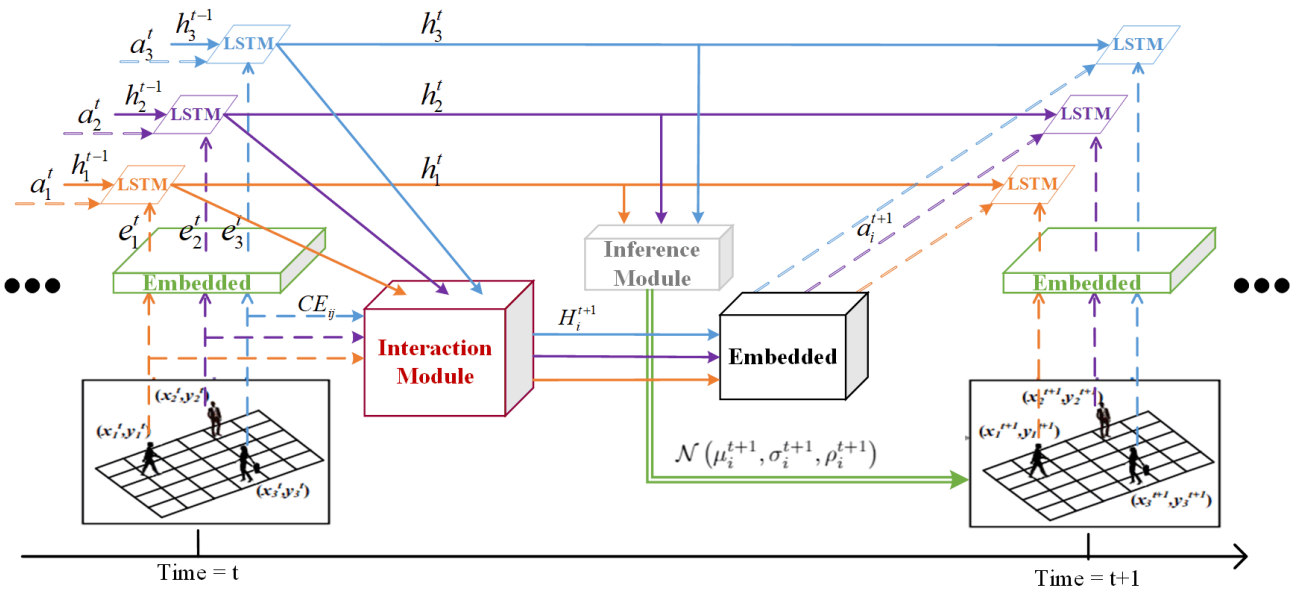



Fig. 3. Overview of our model.

crowd. The correntropy mechanism has the ability to represent everyone's personal space and measure the relative importance of the human-human interactions. Moreover, to effectively incorporate the accurate human-human interaction into the trajectory prediction model, we proposed Interaction Module for autonomously sharing these interactions among all pedestrians in the scene.

## III. Methodology

Pedestrians moving in the scene will adapt their behaviors based on others, and different interactions among them will cause different responses. Motivated by this, we propose a model which can understand human-human interactions while making predictions of trajectories.

In this section, we provide a detailed introduction of our proposed novel LSTM structure combining with an "Interaction Module" to make predictions of all trajectories in the scene. And we refer to our whole proposed model as the Interaction-Aware LSTM model.

### A. Problem Definition

First, we assume that frames of the video with a fixed interval are preprocessed to obtain the spatial coordinates of each pedestrian. We denote the coordinates $(x_i^t, y_i^t) \in \mathbb{R}^2$ of pedestrian $i$ at time $t$ as $\vec{p}_i^t$. Then, we formally describe the trajectory prediction problem as follow. Predict the future trajectory $\Gamma_i = (\vec{p}_i^{obs+1}, ..., \vec{p}_i^{pred})$ of the target pedestrian $i$ from time steps $t = T_{obs+1}, ..., T_{pred}$, taking account his or her own past trajectory $\mathcal{H}_i = (\vec{p}_i^1, ..., \vec{p}_i^{obs})$ from time steps $t = 1, ..., T_{obs}$ and other pedestrians' trajectory information in the scene $\{\mathcal{H}_j : j \in 1, 2, ..., N, j \neq i\}$, where $N$ denotes the number of pedestrians in the scene.

Our goal is to learn the parameters $W^*$ of a model in order to predict the future locations of each pedestrian between $t = T_{obs+1}$ and $t = T_{pred}$. Formally,

$$\Gamma_i = f(\mathcal{H}_1, \mathcal{H}_2, ..., \mathcal{H}_N; W^*) \tag{8}$$

where $W^*$ is the collection of all parameters used in our model. The details are elaborated in the following section.

### B. Model Description

As mentioned above, trajectory prediction can be considered as a time series prediction problem, and the LSTM network is well suited for this task *i.e.* [20]. In previous work, each vanilla LSTM represents one pedestrian, the spatial coordinates at some point are used as the input, the hidden states of the vanilla LSTM are used to estimate the future locations of pedestrians. In this vanilla LSTM, there is a fatal weakness that each pedestrian is modeled independently by a separate LSTM and the dynamic human-human interactions cannot be shared among pedestrians. The pedestrians are isolated in this framework, which makes them unable to interact with people around them, and the prediction error is large [19]. In order to address this issue, we design a variant structure of LSTM illustrated in Figure 3. In this structure, each LSTM still represents one pedestrian but we introduce an "Interaction Module" for passing or sharing dynamic human-human interactions among pedestrians in the scene. In our framework, each pedestrian is able to receive interaction information from people around him/her.

Firstly, as shown in green "Embedded Module" of Figure 3, at time $t$, the coordinates of the $i^{th}$ pedestrian $(x_i^t, y_i^t)$ are embedded into a vector $\boldsymbol{e}$ as follows:

$$e_i^t = \phi\left(x_i^t, y_i^t; W_e\right) \tag{9}$$

where $\phi$ is the embedding function with ReLU non-linerlity, $W_e$ are the embedding parameters.

Then, in order to share the information with each other, we define an "Interaction Module" trying to extract different relations among pedestrians. As shown in the red "Interaction Module", we construct an "Interaction Tensor" $H_i^t$ as the output of the "Interaction Module" to represent the human-human interactions occurred around the target pedestrian $i$ at time $t$. For sharing the relations, similarly, the "Interaction Tensor" $H_i^t$ is embedded into a vector $\boldsymbol{a}$ as follows.

$$a_i^t = \phi\left(H_t^i; W_a\right) \tag{10}$$

where $\phi$ is the embedding function with ReLU non-linerlity, $W_a$ are the embedding parameters.

The vector $\boldsymbol{e}$ represents the spatial feature information of the target pedestrian and the vector $\boldsymbol{a}$ represents the feature information of human-human interactions acted on the target pedestrian in the scene. Finally, we concatenate them as one input to the LSTM cell, which is formulated as:

$$h_i^t = \text{LSTM}\left(h_i^{t-1}, concat(e_i^t, a_i^t); W_l\right) \tag{11}$$

where $W_l$ are the LSTM parameters and all LSTM parameters are shared across pedestrians at the same time step.

*1) Inference Module:* In inference module, we assume a bivariate Gaussian distribution parameterized by the mean $\mu_i^t = (\mu_x, \mu_y)_i^t$, standard deviation $\sigma_i^t = (\sigma_x, \sigma_y)_i^t$, and correlation coefficient $\rho_i^t$ to estimate the predicted coordinates. These parameters at time $t+1$ are determined by the hidden state $h_i^t$ at time $t$ passing through a linear layer $W_o$ as follows:

$$\left(\mu_i^{t+1}, \sigma_i^{t+1}, \rho_i^{t+1}\right) = W_o h_i^t \tag{12}$$

The predicted coordinates are given by:

$$(\hat{x}_i^{t+1}, \hat{y}_i^{t+1}) \sim \mathcal{N}\left(\mu_i^{t+1}, \sigma_i^{t+1}, \rho_i^{t+1}\right) \tag{13}$$

Our model is jointly trained by minimizing the negative log-likelihood loss $L_i$ ($L_i$ represents the $i^{th}$ trajectory) as follows:

$$L^i\left(W_e, W_a, W_l, W_o\right) = -\sum_{t=T_{obs}+1}^{T_{pred}} \log\left(\mathbb{P}\left(x_i^t, y_i^t | \sigma_i^t, \mu_i^t, \rho_i^t\right)\right) \tag{14}$$

Note that the loss is calculated over the entire trajectories in the training datasets. We jointly back-propagate through our model at every time step and tuning the parameters to minimize the loss.

*2) Interaction Module:* Pedestrians have their own walking motion modes, the velocities are different, the orientations are different, and their goals are always different. Pedestrians respect personal space for each pedestrian. Additionally, pedestrians have the unique ability to take the corresponding action based on the information gathered around them and adjust their behavior according to the real-time interactions with other objects in the scene. As argued above, the human-human interactions are significant in predicting the next move of the pedestrians, some existing methods [8], [17]–[19], [27], [35], [83] believe that the human-human interactions are equally important or make specific settings or simple "repulsion" or "attraction" functions to measure the importance of human-human interactions, which is inappropriate in trajectory prediction case, especially among crowds. For instance, a person who enter the personal space of target pedestrian definitely leads to more interactions than a person who is distant from target pedestrian.

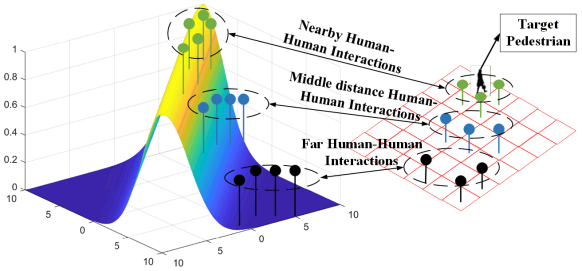



Fig. 4. Calculation the correntropy between the target pedestrian and the pedestrians around him/her.

Note that, as a robust nonlinear similarity measure, the correntropy is naturally suitable in such pedestrians trajectory prediction case. The correntropy has the ability to build personal space for each pedestrian through setting suitable Gaussian kernel. As shown in Figure 4, some pedestrians (green representation) enter the personal space of target pedestrian, which greatly effects the target pedestrian and the generated interactions are obviously huge. Also, the correntropy is very robust to impulsive noises or outlier points thanks to the properties of Gaussian kernel [62], [64]–[67], [84]–[87]. Hence, in this case, human-human interactions caused by the pedestrians who are far away from the target pedestrian, the black representations shown in Figure 4, will be considered outliers and attenuated the effects. Considering the great advantages of correntropy, first, we calculate the correntropy between the target pedestrian and the pedestrians around him/her. The correntropy representing human-human interaction between the $i^{th}$ pedestrian and the $j^{th}$ pedestrian is given by:

$$CE_{ij} = \exp(-\|\vec{p}_i^t - \vec{p}_j^t\|^2 / 2\sigma^2) \tag{15}$$

where $\vec{p}_i^t$ and $\vec{p}_j^t$ are the spatial coordinates $(x^t, y^t)$ of the $i^{th}$ and the $j^{th}$ pedestrian at time $t$, $\|\vec{p}_i^t - \vec{p}_j^t\|^2$ calculates the Euclidean distance between the $i^{th}$ and the $j^{th}$ pedestrian. The value of the correntropy (between 0 and 1) represents the relative importance of their interactions (1 is the most, 0 is the least).

After obtaining the relatively importance of the human-human interactions, we use the obtained relatively importance to construct an "Interaction Tensor", which represents the human-human interactions among crowds in the scene. As the target pedestrian $(x_1, y_1)$ shown in Figure 5, the target pedestrian has his/her own walk state and would be influenced by other pedestrians. In other words, the trajectory of target pedestrian is jointly determined by his/her own walk state and interactions generated by other pedestrians. We first extract the feature representations $h_i^t$ of other pedestrians in the scene. Notice that, in LSTM network, the hidden state $h_i^t$ at time $t$ captures the latent representation of the $i^{th}$ pedestrian at that instant. Then we multiply the value which represents the relative importance of the human-human interaction and the

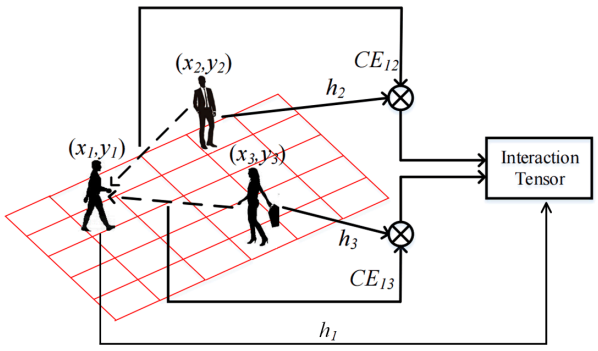



Fig. 5. Diagram of the Interaction Module.

hidden state, to construct an "Interaction Tensor" for sharing feature information among crowds in the scene. More specifically, we construct the "Interaction Tensor" $H_i^t$ of the $i^{th}$ pedestrian including itself hidden state $h_i^{t-1}$ and interactions from other pedestrians at time $t-1$ as follows:

$$H_i^t(x_i, y_i, :) = h_i^{t-1} + \sum_{j=1, j\neq i}^{M_i} CE_{ij} \cdot h_j^{t-1} = \sum_{j=1}^{M_i} CE_{ij} \cdot h_j^{t-1} \tag{16}$$

where $h_j^{t-1}$ is the hidden state of the LSTM representing the $j^{th}$ pedestrian at time $t-1$, and $M_i$ is the set of other pedestrians who have interactions with the target pedestrian $i$ in the coordinate $(x_i, y_i)$.

## IV. Experiments and Analysis

In this section, we demonstrate the experimental results of our approach in two public datasts: ETH [35] and UCY [88]. The ETH datast has total 750 pedestrians and two scenes (ETH and Hotel). The UCY dataset has totoal 786 pedestrians and three scenes (ZARA01, ZARA02, and UCY). These two datasets are both collected from real world, containing complex situations such as pedestrians walking in groups, non-linear trajectories with different velocities, intentionally avoiding collisions and other challenging behaviors, which is suitable for our experiments.

**Evaluation Metrics:** Similar to [3], [19], [27], we use two following metrics. Assume $N$ is the number of the trajectories in the testing process, $\vec{p}_{i,pred}^t$ represents the predicted spatial coordinates $(x^t, y^t)$ of the $i^{th}$ pedestrian at time $t$ and $\vec{p}_{i,obs}^t$ represents the respective observed location.

- *Average Displacement Error:* Introduced in [3], this error calculates the mean distance between all predicted points and the actual points in one trajectory .

$$ADE = \frac{\sum_{i=1}^{N} \sum_{t=T_{obs}+1}^{T_{pred}} \left(\vec{p}_{i,pred}^t - \vec{p}_{i,obs}^t\right)^2}{N\left(T_{pred} - (T_{obs}+1)\right)} \tag{17}$$

- *Final Displacement Error:* Introduced in [19], this error calculates the mean distance between the final predicted point and the final actual point at the end of the prediction process $T_{pred}$.

$$FDE = \frac{\sum_{i=1}^{N} \sqrt{\left(\vec{p}_{i,pred}^t - \vec{p}_{i,obs}^t\right)^2}}{N} \tag{18}$$

**Baselines:** We choose the "Social LSTM" model and the "Social Attention" model as our baselines to compare.

- *Social LSTM Model*: The "Social LSTM" model outperforms the linear model, the "Social Force" model [18] and the Interacting Gaussian Processes model [7].
- *Social Attention Model*: The "Social Attention" model outperforms the "Social LSTM" model on some datasets.

**Implementation Details:** During training, we use a leave-one-out approach where we train and validate our model on 4 sets and test on the remaining one. Our baselines, "Social LSTM" model and "Social Attention" model, are also trained in the same way. During testing, we observe the trajectory for 8 frames and predict the next 12 frames. The frame rate is 0.4, which means $T_{obs} = 3.2secs$, $T_{pred} - T_{obs} = 4.8secs$. It is also the same in our baselines.

The "Interaction Tensor" size is $N_o \times N_o \times D$ , where $N_o$ is 128, $D$ is the dimension of the hidden state and we set $D$ as 128 for all the LSTM models. All the inputs are embedded into a 64 dimensional vector with ReLU nonlinearitym. The batch size is 8 and the model is trained for 150 epochs using Adam with an initial learning rate of 0.001. These settings are the same with baselines. Importantly, we set different values of $\sigma$ to study the performance.

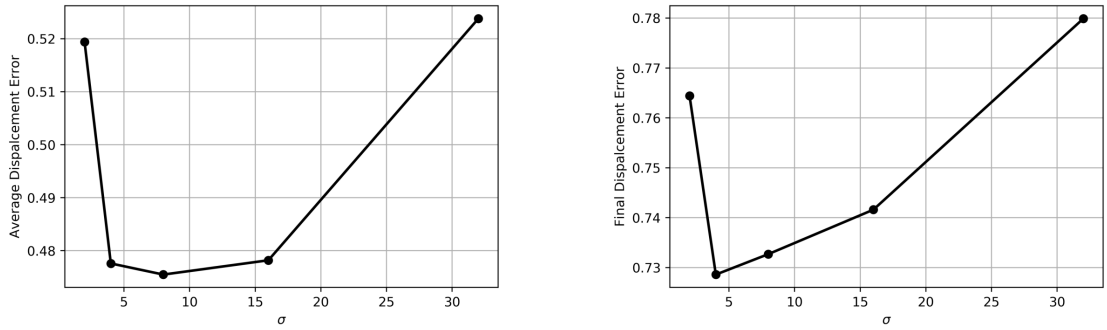
During testing, we use our trained model to determine the parameters of the bivariate Gaussian distribution and then sample from it to obtain the coordinates $(\hat{x}, \hat{y})_i^t$ of the $i^{th}$ pedestrians according to Equation (13). In order to reduce random errors, we run each test for 50 times and calculate the averages and variances as final results.

From time $T_{obs+1}$ to $T_{pred}$, we replace the actual coordinates $(x_i^t, y_i^t)$ in Equation (9), (10) and (11) with the predicted coordinates $(\hat{x}_i^t, \hat{y}_i^t)$ to make predictions. Also, the predicted coordinates are used to calculate the "Interaction Tensor" in Equation (15) and (16).

The whole model is trained on a single Titan-X GPU with a PyTorch implementation and all the results as well as variances are shown in Table II.

### A. Performance Study of $\sigma$

Mentioned in Equation 2, the parameter $\sigma$ denotes the Gaussian kernel size, namely the bandwidth which is related to personal space. According to [89], [90], Gaussian kernel size has the great anti-jamming ability for noise in data, and the data is almost useless beyond the $3\times\sigma$ range (99.73%). In this case, the value of $\sigma$ can represent the radius of one domain where the human-human interactions are effective to a certain extent. In other words, the value of $\sigma$ determines the amount of domain information the target pedestrian can receive in the scene. If the interaction occurred out the effective domain, it would be considered outlier and excluded from our proposed mechanism. The Gaussian kernel is also able to represent personal space in which the generated interaction is huge. The Gaussian kernel has strong nonlinearity and has a large value in a certain range, which is similar to the personal space of a pedestrian. We conduct experiments to show how different values of $\sigma$ affect the performance of our model. The values




(a) ADE (b) FDE

Fig. 6. Results of using different $\sigma$.

TABLE I
EVALUATION RESULTS ON ALL THE DATASETS.

| | ETH | Hotel | ZARA01 | ZARA02 | UCY |
|---|---|---|---|---|---|
| Number of pedestrians | 360 | 390 | 148 | 204 | 434 |
| Number of frames | 8603 | 11401 | 1088 | 877 | 1352 |
| Number of frames within over one pedestrian | 7737 | 10516 | 1075 | 862 | 1352 |

TABLE II
EXPERIMENTAL RESULTS ON ALL DATASETS.

| Metric | Dataset | Social LSTM/*Var* | Social Attention/*Var* | Ours ($\sigma = 2$)/*Var* | Ours ($\sigma = 4$)/*Var* | Ours ($\sigma = 8$)/*Var* | Ours ($\sigma = 16$)/*Var* | Ours ($\sigma = 32$)/*Var* |
|---|---|---|---|---|---|---|---|---|
| Average Displacement Error (ADE) | ETH | 0.5045 */0.0001* | 0.4834 */0.0001* | 0.4926 */0.0001* | **0.4285** */0.0001* | 0.4321 */0.0001* | 0.4970 */0.0001* | 0.4902 */0.0001* |
| | Hotel | 0.5293 */0.0002* | 0.5423 */0.0002* | 0.6389 */0.0002* | **0.5044** */0.0001* | 0.5536 */0.0001* | 0.5136 */0.0002* | 0.5731 */0.0001* |
| | ZARA01 | 0.5404 */0.0032* | 0.6479 */0.0041* | 0.4930 */0.0046* | 0.6217 */0.0036* | 0.4565 */0.0033* | **0.4382** */0.0019* | 0.4474 */0.0008* |
| | ZARA02 | 0.4827 */0.0012* | **0.2336** */0.0002* | 0.4025 */0.0010* | 0.3639 */0.0007* | 0.4159 */0.0009* | 0.4002 */0.0013* | 0.5664 */0.0016* |
| | UCY | 0.5653 */0.0001* | 0.5048 */0.0001* | 0.5700 */0.0016* | **0.4695** */0.0001* | 0.5196 */0.0001* | 0.5422 */0.0001* | 0.5421 */0.0001* |
| | *Average* | 0.5244 */0.00096* | 0.4824 */0.00096* | 0.5194 */0.0015* | 0.4776 */0.00092* | **0.4755** */0.0009* | 0.4782 */0.00072* | 0.5238 */0.00054* |
| Final Displacement Error (FDE) | ETH | 0.9477 */0.0003* | 0.9581 */0.0011* | 0.9213 */0.0005* | **0.7679** */0.0004* | 0.8522 */0.0003* | 0.9297 */0.0007* | 0.8608 */0.0004* |
| | Hotel | 0.9362 */0.0011* | 0.8709 */0.0009* | 0.9635 */0.0011* | **0.7975** */0.0006* | 0.9377 */0.0010* | 0.8734 */0.0013* | 0.9489 */0.0007* |
| | ZARA01 | 0.9717 */0.0271* | 1.0644 */0.0298* | 0.5433 */0.0316* | 0.8369 */0.0310* | 0.5489 */0.0414* | 0.4604 */0.0115* | **0.4213** */0.0091* |
| | ZARA02 | 0.7099 */0.0050* | **0.2486** */0.0005* | 0.5334 */0.0051* | 0.5477 */0.0031* | 0.5574 */0.0028* | 0.5685 */0.0047* | 0.8710 */0.0052* |
| | UCY | 0.8546 */0.0004* | 0.7289 */0.0004* | 0.8606 */0.0006* | **0.6929** */0.0004* | 0.7674 */0.0008* | 0.8761 */0.0007* | 0.7978 */0.0006* |
| | *Average* | 0.8840 */0.00678* | 0.7742 */0.00654* | 0.7644 */0.00778* | **0.7286** */0.00710* | 0.7327 */0.00926* | 0.7416 */0.00378* | 0.7799 */0.0032* |

of $\sigma$ are selected from 2, 4, 8, 16, and 32, which covers a reasonable range of integer values.

We run experiments on the datasets mentioned above and averaged out the two metrics to plot the performance change in Figure 6. Tabel II illustrates the results with different values of $\sigma$. It can be seen that Figure 6a and 6b have the same trends. When $\sigma = 2$, the error of our model is large because little information around the target pedestrian would be shared, which means the target pedestrian barely receive the human-human interactions around him/her in the scene. When $\sigma$ is too large, the performance decreases. A possible explanation is that if $\sigma$ is larger than a person's receptive field, useless information such as a person who is so far away from the target pedestrian, would be considered to make predictions. According to Figure 6a and 6b, our model achieves the minimum average displacement error when choosing $\sigma = 8$ and the minimum final displacement error when choosing $\sigma = 4$. When $\sigma$ is appropriate (between 4 to 16), our model can achieve competitive results.

### *B. Quantitative Analysis*

First, we analyze the complexity of the five scenes and the results are shown in Table I. According to Table I, the numbers of pedestrians in each case have big differences, there are less than 150 pedestrians in case ZARA01 and only 204 pedestrians in case ZARA02, which means these two cases are relatively sparser. We also calculate the number of frames within more than one pedestrian where human-human interactions occur. It is obvious that the ETH and Hotel datatsets are much more crowded than three other cases, which means the human-human interactions are much more complex and difficult to measure.

As shown in Table II, in general, our model outperforms "Social LSTM" model and "Social Attention" model based on two metrics (we take results with $\sigma = 4$ as examples to analyze). In particular, our model significantly outperforms two other models on relatively crowd datasets ETH, Hotel and UCY. Comparing with the "Social LSTM" model, it only performs better on case ZARA01 than our model with $\sigma = 4$ and worse than our model with other values of $\sigma$

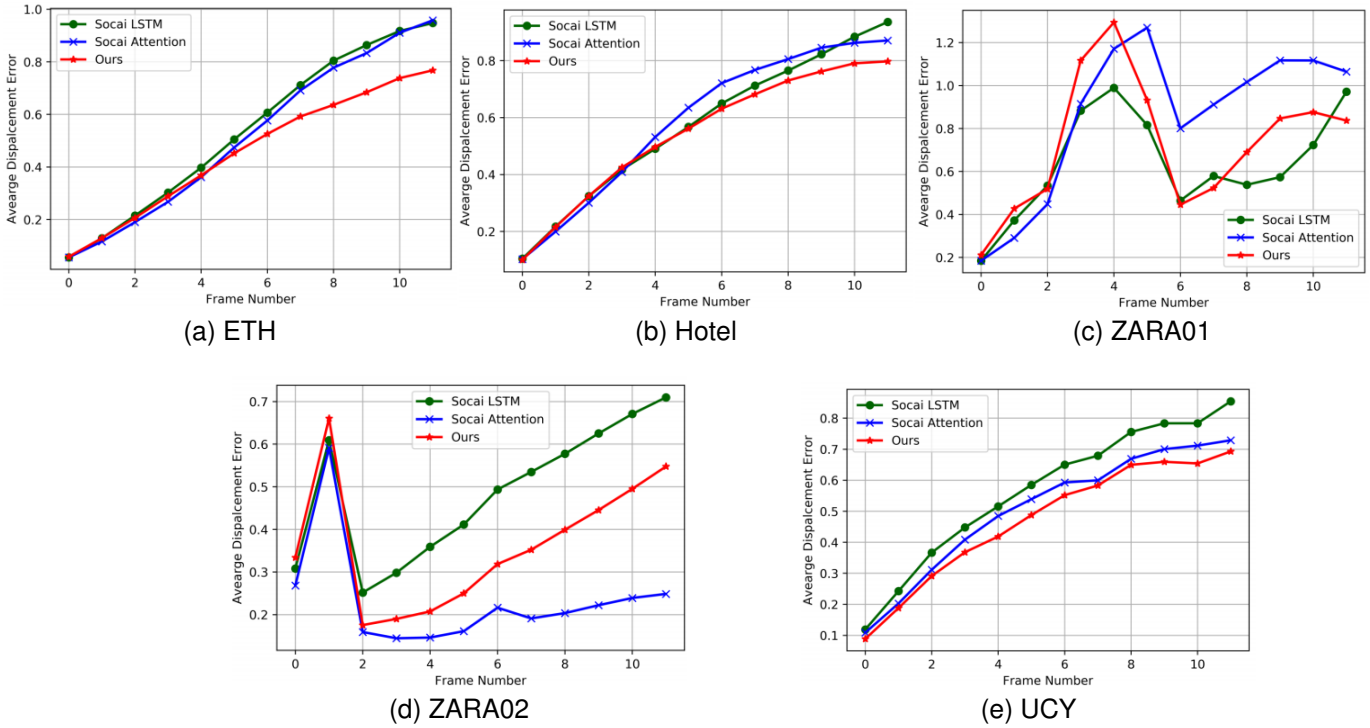



Fig. 7. Variations of the error of each frame. The x-axis represents the frame number to be predicted, ranking from 1 to 12, and the y-axis represents the average displacement error.

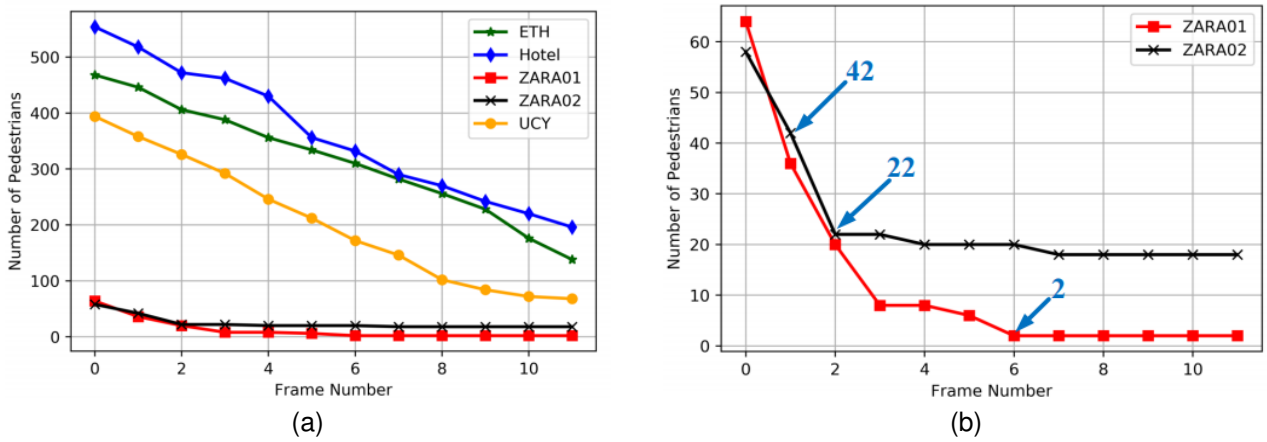



Fig. 8. Variations of the number of pedestrians whose trajectories are predicted.

according to the ADE metric. As the "Social-LSTM" model assumed all the human-human interactions are the same, it cannot balance the different interactions occurred to the same pedestrian and make corresponding predictions, which leads to the highest average prediction ADE and FDE. Comparing with the "Social Attention" model, it achieves the best performance only on case ZARA02, which means the attention mechanism is effective in sparser scenes, but the results prove that our proposed mechanism based on correntropy is more robust to measure the different complex human-human interactions in crowd scene.

Figure 7 plots the change of average error along with the frame number. In Figure 7a, 7b and 7e, the average error increases when frame number increases, it is congenial with reason and common sense. However, in Figure 7c and 7d, the change is dramatic, the average error increases at the beginning and then plunges. In order to find out the reason, we plot the number of pedestrians who participate in trajectory prediction in each frame in Figure 8. We can see from Figure 8a, the number of pedestrians decreases along with the frame in all datasets because the number of pedestrians whose trajectories are longer than 20 frames (8 frames for observation and 12 frames for predictions) is small, especially in relatively sparser case ZARA01 and ZARA02. Figure 8b is an enlarged illustration of number change in scenes ZARA01 and ZARA02. In case ZARA01, there are only 2 pedestrians participating in

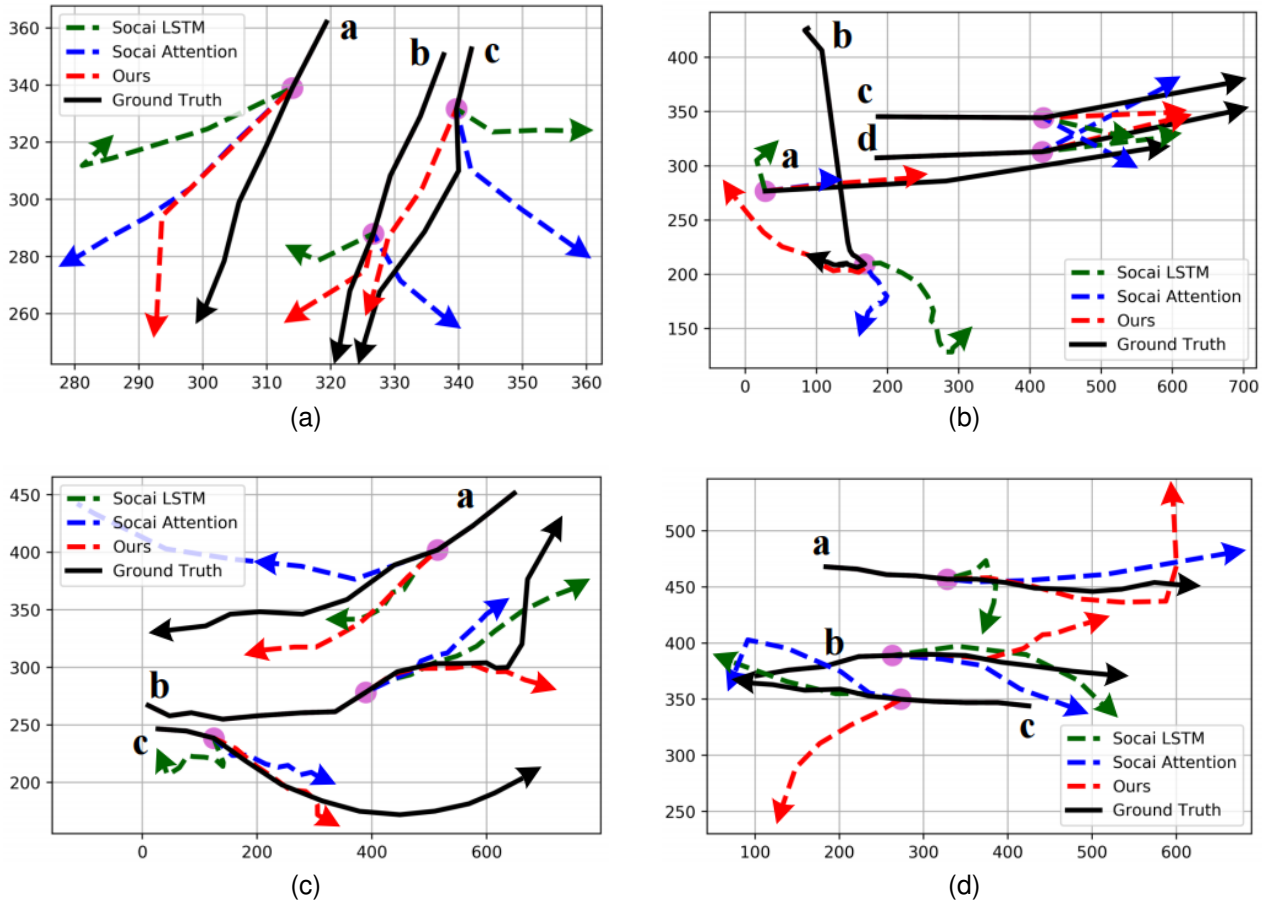



Fig. 9. Illustration of predicted trajectories. These examples contains different complex but common situations in real world.

prediction since frame 6 and this specific person's trajectory decides the whole error. So, the accumulative error before frame 6 disappeared and this situation is corresponding to the sudden drop in Figure 7c. In case ZARA02, from frame 1 to frame 2, the number of pedestrians reduces from 42 to 22. The sudden change of the number greatly decreases the accumulative error and it is a possible reason for the sudden drop in 7d

Above all, the whole experimental results prove that the mechanism we proposed is able to measure the relative importance of the interactions effectively and our model has the ability to understand the different complex human-human interactions in the scene to make predictions.

### C. Qualitative Analysis

In this part, we visualize the predicted trajectories and try to analyze the different interactions in a social way. Some examples of predicted trajectories are shown in Figure 9, the pedestrians are numbered with letters (a, b, c, ...) at the beginning of there trajectories, the purple circle represents the position where the prediction begins. The original data of pedestrians are normalized in above experiments and we remap the data to the real scene in this section. The coordinate axis represents the resolution size of the real scene.

In Figure 9a, the predicted trajectories using our model (red dashed line) are closer to the ground truth (black solid line) than other trajectories generated by "Social LSTM" and "Social Attention" model. In the right side, two pedestrians (pedestrian **b** and **c**) walk in a group and one person (pedestrian **c**) walks alone in the left side. In a social way, as they walk in parallel, they almost have no influence on each other. As the correntropy has the ability to measure the feature similarity, our proposed mechanism attaches relatively small value to the human-human interactions and make predictions. However, the "Social LSTM" model (green dashed line) and the "Social Attention" model (blue dashed line) directly calculate the human-human interactions and make predictions, which makes pedestrians want to enlarge the distance among them and leave each other,namely the deviation phenomena, the trajectory of pedestrian **a** predicted by "Social LSTM" model even turns around at the end.

In Figure 9b, we can see from the vertical trajectory (pedestrian **b**), the direction of the trajectory predicted by our model is totally opposite to the trajectory generated by two other models, because "Social LSTM" model and "Social Attention" model measure the interaction caused by pedestrian **a** in a wrong way. Considering the pedestrians (pedestrian **c** and pedestrian **d**) who are walking in parallel, the phenomena is similar to that in Figure 9a, the "Social LSTM" model and the "Social Attention" model make an incorrect prediction of the future positions based on their assumptions. With our

proposed mechanism, the relative importance of human-human interactions are correctly measured and the predicted results are more accurate.

In Figure 9c, three single persons walk in different directions with different velocity, the interactions between them are different. The results illustrate that our model can measure the interactions more correctly and make better predictions. According to the middle trajectory (pedestrian **b**), our model fails to make predictions of the sudden turn and we will focus on how to make better predictions of "emergencies" such as sudden turn or sudden stop in our future work.

In Figure 9d, in a social point of view, the pedestrians (pedestrian **b** and **c**) walk towards to each other and the distance between them is safe enough according to the ground truth, however, our proposed mechanism is sensitive to the opposite direction according to Equation 15, so our mechanism assigns a high importance to the interaction, which leads to the deviation of prediction. Then, the predicted trajectory of pedestrian **b** leads to a sharp turn of pedestrian **a** to avoid collision. We will take the pose direction into account to solve this kind of mistakes in our future work.

### D. Comparison with Existing Works

We compare our model with several recent existing works: (1) Social-LSTM (S-LSTM) [19]: A method combines the information from all neighboring states by introducing "Social" pooling layers. (2) MX-LSTM [11]: A method capture the interplay between tracklets and vislets to forecast positions and head orientations of an individual. (3) SR-LSTM [12]: A method activates the utilization of the current intention of neighbors and pass the message in the crowd.

As shown in Table III, our model outperforms other methods in relative crowded ETH and UCY dataset, in which the human-human interactions are much more complex and difficult to measure. This indicates that our proposed mechanism based on correntropy is more robust to measure the different complex human-human interactions in crowd scene. In relative sparse ZARA01 and ZARA02 dataset, our model performs better than other models on the FDE metrics, and achieves competitive results with other models on ADE metrics. The results prove that our model also has an effect on sparse scenes.

TABLE III
COMPARISON WITH SEVERAL RECENT MODELS.

| Method | Performance (ADE/FDE) | | | |
|---|---|---|---|---|
| | ETH | UCY | ZARA01 | ZARA02 |
| S-LSTM | 0.50/0.94 | 0.57/0.85 | 0.54/0.97 | 0.48/0.71 |
| MX-LSTM | −/− | 0.49/1.12 | 0.59/1.31 | 0.35/0.79 |
| SR-LSTM | 0.63/1.25 | 0.51/1.10 | **0.41**/0.90 | **0.32**/0.70 |
| OURs | **0.43**/**0.77** | **0.48**/**0.73** | 0.44/**0.46** | 0.36/**0.55** |

## V. CONCLUSION

In this paper, we designed an Interaction-Aware LSTM model to predict future trajectories of pedestrians.

Our model has the ability to measure the relative importance of different human-human interactions occurred in the scene, and then share weighted feature representation through the "Interaction Module" for trajectory prediction. Experiments demonstrate that our proposed model can achieve better performance than several latest methods on two public datasets with five scenes, especially on crowded scenarios. In addition, our qualitative analysis indicates that our model successfully predicts various behaviors arising from human-human interactions in a social way, such as walking as a group or walking in parallel.

As for our future work, we would like to extend our approach to more different scenes and make improvements on our "Interaction Module" for predicting "emergencies" or sudden turns.

## REFERENCES


[1] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," *Computer Graphics Forum*, vol. 26, no. 3, pp. 655–664, 2010.

[2] Y. Luo, P. Cai, A. Bera, D. Hsu, W. S. Lee, and D. Manocha, "Porca: Modeling and planning for autonomous driving among many pedestrians," *IEEE Robotics & Automation Letters*, vol. PP, no. 99, pp. 1–1, 2018.

[3] S. Pellegrini, A. Ess, K. Schindler, and L. J. V. Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *IEEE International Conference on Computer Vision*, 2009, pp. 261–268.

[4] R. P. D. Vivacqua, M. Bertozzi, P. Cerri, F. N. Martins, and R. F. Vassallo, "Self-localization based on visual lane marking maps: An accurate low-cost approach for autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. PP, no. 99, pp. 1–16, 2017.

[5] S. Thrun, M. Bennewitz, W. Burgard, A. B. Cremers, F. Dellaert, D. Fox, D. Hahnel, C. Rosenberg, N. Roy, and J. Schulte, "Minerva: A second-generation museum tour-guide robot," in *IEEE International Conference on Robotics & Automation*, vol. 3, 2002, pp. 1999–2005.

[6] J. J. Leonard and H. F. Durrant-Whyte, "Application of multi-target tracking to sonar-based mobile robot navigation," in *IEEE Conference on Decision & Control*, vol. 29, 1990, pp. 3118–3123.

[7] P. Trautman and A. Krause, "Unfreezing the robot: Navigation in dense, interacting crowds," in *IEEE/RSJ International Conference on Intelligent Robots & Systems*, 2010, pp. 797–803.

[8] M. Luber, J. A. Stork, G. D. Tipaldi, and O. A. Kai, "People tracking with human motion predictions from social forces," in *IEEE International Conference on Robotics & Automation*, 2010, pp. 464–469.

[9] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, "Who are you with and where are you going?" in *CVPR 2011*. IEEE, 2011, pp. 1345–1352.

[10] D. Yang, L. Li, K. Redmill, and Ü. Özgüner, "Top-view trajectories: A pedestrian dataset of vehicle-crowd interaction from controlled experiments and crowded campus," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 899–904.

[11] I. Hasan, F. Setti, T. Tsesmelis, A. Del Bue, F. Galasso, and M. Cristani, "Mx-lstm: mixing tracklets and vislets to jointly forecast trajectories and head poses," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6067–6076.

[12] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng, "Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 085–12 094.

[13] Z. Wan, C. Jiang, M. Fahad, Z. Ni, Y. Guo, and H. He, "Robot-assisted pedestrian regulation based on deep reinforcement learning," *IEEE Transactions on Cybernetics*, vol. 50, no. 4, pp. 1669–1682, 2020.

[14] Y. Feng, Y. Yuan, and X. Lu, "Person reidentification via unsupervised cross-view metric learning," *IEEE Transactions on Cybernetics*, vol. 51, no. 4, pp. 1849–1859, 2021.

[15] N. Shafiee, T. Padir, and E. Elhamifar, "Introvert: Human trajectory prediction via conditional 3d attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 815–16 825.

[16] M. Moussaïd, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz, "The walking behaviour of pedestrian social groups and its impact on crowd dynamics," *PloS one*, vol. 5, no. 4, p. e10047, 2010.

[17] A. Gupta, J. Johnson, L. Feifei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *Computer Vision & Pattern Recognition*, 2018, pp. 2255–2264.

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

[19] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Computer Vision & Pattern Recognition*, 2016, pp. 961–971.

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

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

[22] T. Fernando, S. Denman, A. Mcfadyen, S. Sridharan, and C. Fookes, "Tree memory networks for modelling long-term temporal dependencies," *Neurocomputing*, vol. 304, pp. 64–81, 2018.

[23] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto, "Fast unsupervised ego-action learning for first-person sports videos," in *Computer Vision & Pattern Recognition*, 2011, pp. 3241–3248.

[24] M. S. Ryoo, T. J. Fuchs, L. Xia, J. K. Aggarwal, and L. Matthies, "Early recognition of human activities from first-person videos using onset representations," in *Computer Vision & Pattern Recognition*, 2015.

[25] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised learning of video representations using lstms," in *International Conference on Machine Learning*, 2015, pp. 843–852.

[26] C. Vondrick, H. Pirsiavash, and A. Torralba, "Anticipating the future by watching unlabeled video." *Computer Vision & Pattern Recognition*, 2015.

[27] A. Vemula, K. Muelling, and J. Oh, "Social attention: Modeling attention in human crowds," in *International Conference on Robotics and Automation*, 2018, pp. 1–7.

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

[29] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*. PMLR, 2014, pp. 1764–1772.

[30] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[31] N. Bhargava, S. Chaudhuri, and G. Seetharaman, "Linear cyclic pursuit based prediction of personal space violation in surveillance video," in *2013 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 2013, pp. 1–5.

[32] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan, "Crowded scene analysis: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 3, pp. 367–386, 2015.

[33] A. Sardar, M. Joosse, A. Weiss, and V. Evers, "Don't stand so close to me: Users' attitudinal and behavioral responses to personal space invasion by robots," in *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2012, pp. 229–230.

[34] Z. Bingchen, G. Weimin, Z. Hengyu, C. Huachun, and W. Yanqun, "Research on relationship of passenger's behavior and space," in *2011 IEEE 2nd International Conference on Computing, Control and Industrial Engineering*, vol. 1, 2011, pp. 82–85.

[35] S. Pellegrini, A. Ess, and L. Van Gool, "Improving data association by joint modeling of pedestrian trajectories and groupings," in *European Conference on Computer Vision*, vol. 6311, 2010, pp. 452–465.

[36] N. Schneider and D. M. Gavrila, "Pedestrian path prediction with recursive bayesian filters: A comparative study," in *German Conference on Pattern Recognition*, vol. 8142, 2013, pp. 174–183.

[37] L. Ballan, F. Castaldo, A. Alahi, F. Palmieri, and S. Savarese, "Knowledge transfer for scene-specific motion prediction," in *European Conference on Computer Vision*, 2016, pp. 697–713.

[38] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila, "Context-based pedestrian path prediction," *European Conference on Computer Vision*, vol. 8694, pp. 618–633, 2014.

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

[40] J. Walker, A. Gupta, and M. Hebert, "Patch to the future: Unsupervised visual prediction," in *Computer Vision & Pattern Recognition*, 2014, pp. 2224–2231.

[41] X. Dan, S. Todorovic, and S. C. Zhu, "Inferring "dark matter" and "dark energy" from videos," in *IEEE International Conference on Computer Vision*, 2013, pp. 2224–2231.

[42] Y. Shuai, H. Li, and X. Wang, "Pedestrian behavior understanding and prediction with deep neural networks," in *European Conference on Computer Vision*, 2016, pp. 263–279.

[43] P. Abbeel and A. Y. Ng, "Inverse reinforcement learning," 2016.

[44] A. Y. Ng and S. Russell, "Algorithms for inverse reinforcement learning," in *International Conference on Machine Learning*, vol. 67, no. 2, 2000, pp. 663–670.

[45] B. D. Ziebart, N. D. Ratliff, G. Gallagher, C. Mertz, K. M. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey, and S. S. Srinivasa, "Planning-based prediction for pedestrians," in *IEEE/RSJ International Conference on Intelligent Robots & Systems*, 2009, pp. 3931–3936.

[46] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert, "Activity forecasting," in *European Conference on Computer Vision*, 2012, pp. 201–214.

[47] N. Lee and K. M. Kitani, "Predicting wide receiver trajectories in american football," in *Applications of Computer Vision*, 2016, pp. 1–9.

[48] W. Ma, D. Huang, N. Lee, and K. M. Kitani, "Forecasting interactive dynamics of pedestrians with fictitious play," in *Computer Vision & Pattern Recognition*, 2017, pp. 4636–4644.

[49] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: Properties and applications in non-gaussian signal processing," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5286–5298, 2007.

[50] J. C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*, 1st ed. Springer Publishing Company, Incorporated, 2010.

[51] I. Santamaría, P. P. Pokharel, and J. C. Principe, "Generalized correlation function: Definition, properties, and application to blind equalization," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2187–2197, 2006.

[52] J. Principe, *Information-Theoretic Learning*, 2010.

[53] W. Shi, Y. Gong, X. Tao, and N. Zheng, "Training dcnn by combining max-margin, max-correlation objectives, and correntropy loss for multi-label image classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 7, pp. 2896–2908, 2018.

[54] A. Singh, R. Pokharel, and J. Principe, "The c-loss function for pattern classification," *Pattern Recognition*, vol. 47, no. 1, pp. 441–453, 2014.

[55] N. M. Syed, J. Principe, and P. Pardalos, "Correntropy in data classification," *Springer Proceedings in Mathematics and Statistics*, vol. 20, pp. 81–117, 01 2012.

[56] X. Chen, Y. Jian, J. Liang, and Q. Ye, "Recursive robust least squares support vector regression based on maximum correntropy criterion," *Neurocomputing*, vol. 97, no. Complete, pp. 63–73, 2012.

[57] Y. Feng, X. Huang, S. Lei, Y. Yang, and J. A. K. Suykens, "Learning with the maximum correntropy criterion induced losses for regression," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 993–1034, 2015.

[58] L. Chen, Q. Hua, J. Zhao, B. Chen, and J. C. Principe, "Efficient and robust deep learning with correntropy-induced loss function," *Neural Computing & Applications*, vol. 27, no. 4, pp. 1019–1031, 2016.

[59] J. Xu and J. C. Principe, "A pitch detector based on a generalized correlation function," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1420–1432, 2008.

[60] B. Chen, X. Lei, J. Liang, N. Zheng, and J. C. Principe, "Steady-state mean-square error analysis for adaptive filtering under the maximum correntropy criterion," *IEEE Signal Processing Letters*, vol. 21, no. 7, pp. 880–884, 2014.

[61] S. Zhao, B. Chen, and J. C. Principe, "Kernel adaptive filtering with maximum correntropy criterion," in *International Joint Conference on Neural Networks*, 2011, pp. 2012–2017.

[62] B. Chen, L. Xing, X. Wang, J. Qin, and N. Zheng, "Robust learning with kernel mean $p$-power error loss," *IEEE Transactions on Cybernetics*, vol. PP, no. 99, pp. 1–13, 2018.

[63] H. Ran, H. Bao-Gang, Z. Wei-Shi, and K. Xiang-Wei, "Robust principal component analysis based on maximum correntropy criterion," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1485–1494, 2011.

[64] B. Chen, Y. Xie, X. Wang, Z. Yuan, P. Ren, and J. Qin, "Multikernel correntropy for robust learning," *IEEE Transactions on Cybernetics*, 2021.

[65] B. Du, T. Xinyao, Z. Wang, L. Zhang, and D. Tao, "Robust graph-based semisupervised learning for noisy labeled data via maximum correntropy criterion," *IEEE Transactions on Cybernetics*, vol. 49, no. 4, pp. 1440–1453, 2019.

[66] Y. Wang, Y. Y. Tang, and L. Li, "Correntropy matching pursuit with application to robust digit and face recognition," *IEEE Transactions on Cybernetics*, vol. 47, no. 6, pp. 1354–1366, 2017.

[67] K. Xiong, H. H. C. Iu, and S. Wang, "Kernel correntropy conjugate gradient algorithms based on half-quadratic optimization," *IEEE Transactions on Cybernetics*, vol. 51, no. 11, pp. 5497–5510, 2021.

[68] B. Müller, J. Reinhardt, and M. T. Strickland, *BTT: Back-Propagation Through Time*, 1995.

[69] R. J. Williams and D. Zipser, "Experimental analysis of the real-time recurrent learning algorithm," *Connection Science*, vol. 1, no. 1, pp. 87–111, 1989.
[70] T. Mikolov, M. Karafiat, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," pp. 1045–1048, 2010.
[71] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent nn: First results," *ArXiv: Neural and Evolutionary Computing*, 2014.
[72] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," *Neural Information Processing Systems*, pp. 2980–2988, 2015.
[73] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013, pp. 6645–6649.
[74] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing." *ArXiv: Computation and Language*, 2017.
[75] S. Z. Tajalli, A. Kavousi-Fard, M. Mardaneh, A. Khosravi, and R. Razavi-Far, "Uncertainty-aware management of smart grids using cloud-based lstm-prediction interval," *IEEE Transactions on Cybernetics*, pp. 1–14, 2021.
[76] X. Xu and M. Yoneda, "Multitask air-quality prediction based on lstm-autoencoder model," *IEEE Transactions on Cybernetics*, vol. 51, no. 5, pp. 2577–2586, 2021.
[77] Y. Han, W. Qi, N. Ding, and Z. Geng, "Short-time wavelet entropy integrating improved lstm for fault diagnosis of modular multilevel converter," *IEEE Transactions on Cybernetics*, pp. 1–9, 2021.
[78] A. V. Den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *International Conference on Machine Learning*, 2016, pp. 1747–1756.
[79] A. Karpathy and F. F. Li, "Deep visual-semantic alignments for generating image descriptions," in *Computer Vision & Pattern Recognition*, 2015, pp. 3128–3137.
[80] M. Liu, H. Hu, L. Li, Y. Yu, and W. Guan, "Chinese image caption generation via visual attention and topic modeling," *IEEE Transactions on Cybernetics*, pp. 1–11, 2020.
[81] O. Wu, T. Yang, M. Li, and M. Li, "Two-level lstm for sentiment analysis with lexicon embedding and polar flipping," *IEEE Transactions on Cybernetics*, pp. 1–13, 2020.
[82] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *neural information processing systems*, pp. 3104–3112, 2014.
[83] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2009, pp. 935–942.
[84] Y. Zhu, H. Zhao, X. Zeng, and B. Chen, "Robust generalized maximum correntropy criterion algorithms for active noise control," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1282–1292, 2020.
[85] L. Gao, X. Li, D. Bi, L. Peng, X. Xie, and Y. Xie, "Robust tensor recovery in impulsive noise based on correntropy and hybrid tensor sparsity," *IEEE Transactions on Circuits and Systems II: Express Briefs*, pp. 1–1, 2021.
[86] H. Zhao, D. Liu, and S. Lv, "Robust maximum correntropy criterion subband adaptive filter algorithm for impulsive noise and noisy input," *IEEE Transactions on Circuits and Systems II: Express Briefs*, pp. 1–1, 2021.
[87] W. Shi and Y. Li, "A shrinkage correntropy based algorithm under impulsive noise environment," in *2019 6th International Conference on Information, Cybernetics, and Computational Social Systems (ICCSS)*, 2019, pp. 244–247.
[88] L. Lealtaixe, M. Fenzi, A. Kuznetsova, B. Rosenhahn, and S. Savarese, "Learning an image-based motion context for multiple people tracking," in *Computer Vision & Pattern Recognition*, 2014, pp. 3542–3549.
[89] J. Kivinen, A. J. Smola, and R. C. Williamson, "Online learning with kernels," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2165–2176, 2004.
[90] C. K. I. Williams, "Learning with kernels: Support vector machines, regularization, optimization, and beyond," *Publications of the American Statistical Association*, vol. 98, no. 462, pp. 489–489, 2003.

**Jing Yang** received the B.S. and M.S. degrees in control science and engineering and the Ph.D. degree in pattern recognition and intelligent systems from Xi'an Jiaotong University, China, in 1999, 2002, and 2010, respectively.

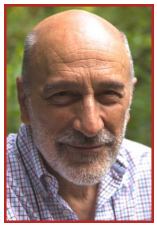
From 1999 to 2003, she was a Research Assistant with the Institute of Automation, Xi'an Jiaotong University. Since 2019, she has been an Associate Professor with the Department of Automation Science and Technology, Xi'an Jiaotong University. Her research interests include machine learning, reinforcement learning, and information theory and their applications to intelligent systems such as autonomous vehicles. Since 2004, she has been a member of Intelligent Vehicles Team, Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University.

**Yuehai Chen** received the B.S. degree in automation from Xi'an Jiaotong University, Xi'an, Shaanxi, China, in 2020, where he is currently pursuing the master's degree in control science and engineering. His research interests include machine learning, artificial intelligence, computer vision, and their applications to intelligent systems.

**Shaoyi Du** received the B.S. degrees in computational mathematics and in computer science, the M.S. degree in applied mathematics, and the Ph.D. degree in pattern recognition and intelligence system from Xi'an Jiaotong University, China, in 2002, 2005, and 2009 respectively.

He was a Post-Doctoral Fellow with Xi'an Jiaotong University from 2009 to 2011 and was with The University of North Carolina at Chapel Hill from 2013 to 2014. He is currently a Professor with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University. His research interests include computer vision, machine learning, and pattern recognition

**Badong Chen** received the B.S. and M.S. degrees in control theory and engineering from Chongqing University, in 1997 and 2003, respectively, and the Ph.D. degree in computer science and technology from Tsinghua University in 2008. He was a Postdoctoral Researcher with Tsinghua University from 2008 to 2010, and a Postdoctoral Associate at the University of Florida Computational NeuroEngineering Laboratory (CNEL) during the period October, 2010 to September, 2012. During July to August 2015, he visited the Nanyang Technological University (NTU) as a visiting research scientist. He also served as a senior research fellow with The Hong Kong Polytechnic University from August to November in 2017. Currently he is a professor at the Institute of Artificial Intelligence and Robotics (IAIR), Xi'an Jiaotong University. His research interests are in signal processing, information theory, machine learning, and their applications to cognitive science and neural engineering. He has published 2 books, 4 chapters, and over 200 papers in various journals and conference proceedings. Dr. Chen is an IEEE Senior Member, a Technical Committee Member of IEEE SPS Machine Learning for Signal Processing (MLSP) and IEEE CIS Cognitive and Developmental Systems (CDS), and an associate editor of IEEE Transactions on Cognitive and Developmental Systems, IEEE Transactions on Neural Networks and Learning Systems and Journal of The Franklin Institute, and has been on the editorial board of Entropy.



**Jose C. Principe** received the B.S. degree from the University of Porto, Portugal, in 1972 and the M.Sc. and Ph.D. degrees from the University of Florida in 1974 and 1979, respectively. He is a Distinguished Professor of Electrical and Biomedical Engineering with the University of Florida, Gainesville, where he teaches advanced signal processing and artificial neural networks (ANNs) modeling. He is a Bell-South Professor and Founder and Director of the University of Florida Computational NeuroEngineering Laboratory (CNEL). He is involved in biomedical signal processing, in particular, the electroencephalogram (EEG) and the modeling and applications of adaptive systems. He has more than 129 publications in refereed journals, 15 book chapters, and over 300 conference papers. He has directed more than 50 Ph.D.dissertations and 61 master's degree theses.

Dr. Principe is Editor-in-Chief of the IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, President of the International Neural Network Society, and formal Secretary of the Technical Committee on Neural Networks of the IEEE Signal Processing Society. He is an AIMBE Fellow and a recipient of the IEEE Engineering in Medicine and Biology Society Career Service Award. He is also a member of the Scientific Board of the Food and Drug Administration, and a member of the Advisory Board of the McKnight Brain Institute at the University of Florida.