Neurocomputing 490 (2022) 258–268

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

SRAI-LSTM: A Social Relation Attention-based Interaction-aware LSTM for human trajectory prediction

Yusheng Peng [a], Gaofeng Zhang [b], Jun Shi [b], Benzhu Xu [b], Liping Zheng [a,b,c,*]

[a] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230061, China
[b] School of Software, Hefei University of Technology, Hefei 230061, China
[c] Anhui Province Key Laboratory of Industry Safety and Emergency Technology, Hefei University of Technology, Hefei 230061, China

ARTICLE INFO

Article history:
Received 22 June 2021
Revised 21 September 2021
Accepted 27 November 2021
Available online 3 December 2021

Keywords:
Trajectory prediction
Social interaction
Social relation
Social relation attention

ABSTRACT

Pedestrian trajectory prediction is one of the important research topics in the field of computer vision and a key technology of autonomous driving system. Walking in groups is a common social behavior in which pedestrians pay more attention to the movements of their companions while walking. Motivated by this idea, we propose a Social Relation Attention-based Interaction-aware LSTM (SRAI-LSTM) to model this social behavior for trajectory prediction. We design a social relation encoder module to capture social relation feature between pedestrians through their relative positions. Afterwards, the social relation features are adopted to acquire social relation attentions among pedestrians. Social interaction modeling is achieved by utilizing social relation attentions to aggregate motion features from neighbor pedestrians. Experimental results on two public pedestrian trajectory datasets (ETH and UCY) demonstrate that our proposed model achieves superior performances compared with state-of-the-art methods on ADE and FDE metrics.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

Pedestrian trajectory prediction is an important research topic in the field of computer vision, and supports various meaningful applications, such as intelligent surveillance system [1,2], autonomous driving system [3,4] and robot navigation system [5,6], etc. This task is challenging because human–human interactions are multi-modal and extremely hard to capture.

To capture these intricate and subtle interactions, the handcrafted energy functions adopted in earlier works [7–9], which require significant feature engineering effort and normally fail to build crowd interactions in crowded spaces [10]. With the rapid development of deep learning, the methods with deep neural networks have significantly improved the performance of trajectory prediction. In these methods, pooling mechanisms [11–13] and graph neural networks mechanisms [10,14,15] are widely used to model social interactions among pedestrians, and all of them have achieved good performance. Furthermore, the different impacts among pedestrians need to be taken into account in social interaction modeling. Inspired by this idea, the attention mechanism is widely adopted to capture the different impacts of pedestrians. In earlier methods [16,17], the attention between pedestrians is captured from their motion features. And in later works, the attention is captured by coupling additional information, such as relative position [18,19], bearing angle [13,20], etc.

The goal of pedestrian trajectory prediction is to predict the socially acceptable future trajectories. As the people meeting scenario shows in Fig. 1, A and B travel in pairs, and C travel alone. The future trajectories in (a) are socially acceptable because C does not break the companion walking behavior of A and B when he moves. The behavior of C in figure (b) who walks between A and B is socially unacceptable because it breaks the social rules. Therefore, it is necessary to pay extra attention to the social relation between pedestrians in social interaction modeling.

Motivated by this insight, a Social Relation Attention-based Interaction-aware LSTM (SRAI-LSTM) is proposed to predict socially acceptable trajectories. In our proposed SRAI-LSTM model, the social relation encoder module is designed to extract social relation features, and the social interactions among pedestrians are modeled through the social interaction module. Specifically, a LSTM is utilized as social relation encoder to continuously learn and update the social relation feature between each pair of pedestrians from their relative positions. In the social interaction module, we introduce a novel attention named social relation attention which is acquired from social relation features and

* Corresponding author at: School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230061, China.
E-mail address: zhenglp@hfut.edu.cn (L. Zheng).

https://doi.org/10.1016/j.neucom.2021.11.089
0925-2312/© 2021 Elsevier B.V. All rights reserved.

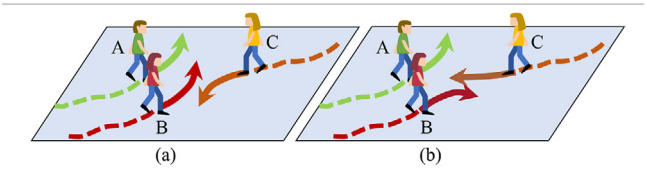


Fig. 1. The figures show a meeting scenario among pedestrians A, B, and C, in which A and B walk in pairs. Our goal is to predict the socially acceptable future trajectories (a), rather than the future trajectories (b) which break social rules.

motion features. And then the social interactions are modeling by aggregating motion features via social relation attentions.

The main contributions of this paper are summarized as follows:

- We propose learning the social relation feature between pedestrians from their relative movements. To our knowledge, this idea has never been mentioned in the research of human trajectory prediction.
- A novel attention mechanism is used for social interaction modeling, in which the attention is acquired by coupling the social relation feature with the latent motion features. Compared with existing attention mechanisms, this is the first one which focuses on social relations among pedestrians.
- Experiments on ETH and UCY datasets show that SRAI-LSTM significantly improves pedestrian trajectory prediction, achieving state-of-the-art performance on two popular benchmarks.

The rest of the paper is organized as follows. A brief overview of related works is provided in Section 2. The proposed SRAI-LSTM model for human trajectory prediction is described in Section 3. Experimental comparisons with state-of-the-art methods on the ETH [21] and UCY [22] pedestrian trajectory datasets are presented and discussed in Section 4. Finally, the conclusions and future works are provided in Section 5.

2. Related work

2.1. Human trajectory prediction

Forecasting human trajectory has been researched for decades. In the early stages, many classic approaches are widely applied, such as linear regression and Kalman filter [23], Gaussian processes [24] and Markov decision processing [25]. However, it is hard to model complex social interactions and normally fail in crowded scenes via these methods.

In recent years, Long Short-Term Memory (LSTM) models have achieved great success in various sequence prediction tasks like activity recognition [37,38] and motion prediction [39], which are widely used in pedestrian trajectory prediction methods [11,26,18,27]. To predict multiple socially acceptable trajectories, Generative Adversarial Networks (GANs) are introduced to generate a socially compliant set of possible trajectories for each pedestrian, among which the typical methods include Social-GAN [12], GD-GAN [28], Social-Ways [13], Sophie [29], Goal-GAN [30], etc. Conditional Variational Auto-Encoder (CVAE) as another popular generative model are used in various trajectory prediction methods [31,32]. As the Temporal Convolutional Network (TCN) reached or even exceeded the Recurrent Neural Network (RNN) in multiple tasks, some scholars use TCN to replace the RNN model and achieve success in pedestrian trajectory prediction [33,34]. With the success of the Transformer model using the self-attention mechanism in the field of natural language processing and computer vision, Yu et al. [35] and Ye et al. [36] introduces Transformer to trajectory prediction and achieve state-of-the-art performance.

In view of the recurrent property of LSTM and its success in sequence task and trajectory prediction, we adopt LSTM as the backbone to build a novel recurrent net for trajectory prediction. Besides, an extra LSTM is employed as an encoder to capture social relation features between pedestrians.

2.2. Social interaction modeling

Handcrafted rules and energy parameters [7,22,40] have been used to capture social interactions but fail to generalize properly. In some recent approaches [11,12,41], pooling mechanisms have been used to model social interactions among pedestrians in local or global neighborhoods. With the development of Graph Neural Network, Graph Convolutional Networks (GCN) and Graph Attention Networks (GAT) are widely used to model spatial–temporal interaction in Seq2Seq trajectory prediction models [10,14,15]. Moreover, Zhang et al. [18] and Hu et al. [42] propose to adopt message passing mechanisms to model social interactions.

In the view that pedestrians have different impacts on social interactions, Fernando et al. [16] propose to acquire the attention of pedestrians from their latent motion features. However, Amirian et al. [13] captured attention by coupling spatial features, such as Euclidean distance and bearing angle between pedestrians. Besides, various variations of the attention mechanism have been widely adopted in recent methods [31,42,43] of trajectory prediction. Spatial information and motion information are always utilized to capture attention, but the social relations between pedestrians are ignored in the mentioned methods.

In this paper, we adopt an attention-based aggregation function to model social interactions. To model the different impacts of neighbor pedestrians on social interaction, we couple the latent motion features and social relation features to capture the attention, and the social interaction feature is acquired by attentively aggregating the motion features of neighbor pedestrians.

2.3. Social relation modeling

Social relation is the close association between individual persons and forms the basic structure of our society. In Location-aware Social Networks (LSNs), the spatio-temporal trajectories are employed to learn social relations for location prediction [44], check-in time prediction [45], and human mobility prediction [46], etc. And beyond that, recognizing social relations between persons can empower intelligent agents to better understand the behaviors or emotions of human beings [47] in intelligent systems.

In human trajectory prediction, Sun et al. [48] propose to consider the relation between pedestrians in social interaction modeling, in which the social relation is directly annotated by relation label. In the annotations, using 0/1 to represent whether two pedestrians are in the same group. And the view of pedestrian grouping has also been used in [49,50].

In this paper, we propose adopting a neural network to learn the features of social relations between pedestrians. Specifically, we adopt an LSTM model to model the spatio-temporal correlation of pedestrians' location to capture the social relation feature between pedestrians. The learned social relation feature is employed to calculate attention in the social interaction module.

3. Proposed method

The overview of the SRAI-LSTM framework is illustrated in Fig. 2. The social relation encoder is designed to model the temporal correlation from the relative positions between pedestrians to

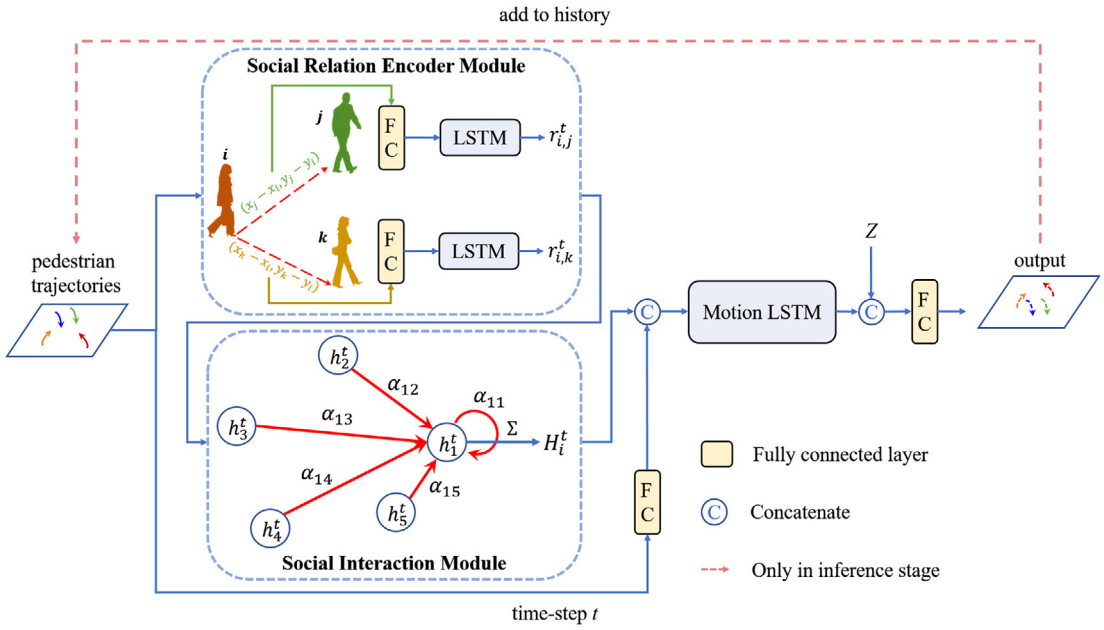


Fig. 2. The architecture of the proposed SRAI-LSTM model. SRAI-LSTM adopts a recurrent structure. In SRAI-LSTM, a motion LSTM is designed to capture latent motion features for each pedestrian. At each time-step, the social relation feature between each pair of pedestrians is captured by the social relation encoder module. And social interaction is modeled by aggregating motion features via social relation attention in social interaction module. The latent code Z is used to generate multiple future trajectories in a stochastic version. In particular, in the inference stage, the predicted coordinates are added to the historical trajectory to infer the next future positions.

acquire social relation features. And the social interactions are modeled by a social interaction module via social relation attention. The embedding of position embedding layer and the social interaction context output from social interaction module are treated as inputs to capture latent motion feature by a LSTM.

3.1. Problem formulation

In this paper, we address the problem of pedestrian trajectory prediction in the crowd scenes. For better modeling the social interactions among pedestrians, we focus on two-dimensional coordinations of pedestrians in the world coordinate system at specific key time-steps. For each sample, we assume there are N pedestrians involved in the scene. Given certain observed positions $\{p_i^t|(x_i^t,y_i^t), t=1,2,\ldots,T_{obs}\}$ of pedestrians i of T_{obs} time-steps, our goal is predicting the positions $\{p_i^{t'}|(\widehat{x}_i^{t'},\widehat{y}_i^{t'}), t'=T_{obs}+1,T_{obs}+2,\ldots,T_{pred}\}$ of future T_{pred} time-steps.

3.2. Social relation encoder module

In crowd scenes, pedestrians with various social relations exhibit different social behaviors. For instance, people who walk in pairs have the same motion patterns and remain nearly the same distance apart. Not only that, they tend to pay higher attention to their companions' behavior when making motion decisions. Therefore, the social relation between pedestrians is an important factor that cannot be ignored in the human trajectory prediction. To our knowledge, this view has never been mentioned in existing works. In this section, we design a model to learn the latent social relation of such behavior of walking in pairs between pedestrians from their spatio-temporal trajectories.

As mentioned above, the continuous and steady distances between pedestrians is a conspicuous manifestation of the behavior of walking in pairs. Therefore, we choose to capture the social relation feature between pedestrians from their relative positions via a spatio-temporal model. Concretely, for each pair of pedestrians, an LSTM is used to encode the relative positions recurrently to capture the social relation feature. As illustrated in Fig. 3, for the target pedestrian 1, the relative positions of the neighbor pedestrians are calculated and served as the current input of the LSTM encoder. The current hidden state of the LSTM encoder represents the feature of the social relation between the pair of pedestrians. We term this LSTM as R-LSTM (LSTM for social relation encoding):

$$e_{ij}^t = \phi_r\left(x_j^t - x_i^t, y_j^t - y_i^t; W_{re}\right) \tag{1}$$

$$r_{ij}^t = \mathrm{R-LSTM}\left(r_{ij}^{t-1}, e_{ij}^t; W_r\right) \tag{2}$$

where e_{ij}^t is embedding vector of the relative position. W_{re} denotes the parameter for the embedding function ϕ_r. r_{ij}^t is the hidden state of the R-LSTM which represent the learned social relation feature between pedestrian i and j at the time-step t, W_r is the R-LSTM weight and is shared among all the sequences.

As time-steps increase, the learned features of social relations will become more believable. Social relation features are used to capture social relation attention in social interaction module.

3.3. Social interaction module

Vanilla LSTM used for per person does not capture human–human interactions. To model the social interaction behavior between pedestrians, pooling mechanisms [11,49,50] are adopted to aggregate hidden states among pedestrians on occupancy maps. Recently, the trajectory prediction models [18,13,43] that adopt attention mechanisms to model social interaction have achieved good performance. Because the attention mechanism can help aggregating shared information by paying different attention to neighboring pedestrians. Thus, we adopt an attention-based aggregation function to model social interactions.

Similar to [16,28,18], the latent motion features are used to capture the attention of pedestrians. In this section, we introduce a novel attention called social relation attention which is captured by coupling pedestrians' latent motion features and the social rela-

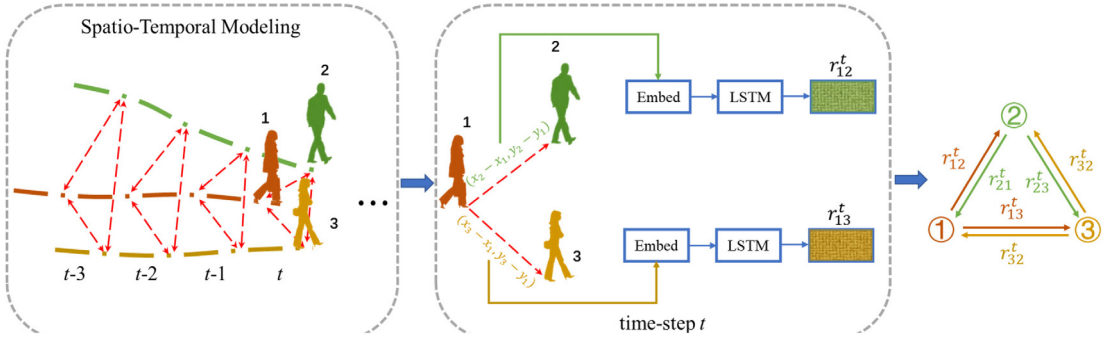


Fig. 3. We model the spatio-temporal correlation of pedestrians' positions to capture the social relation features between each pair of pedestrians. For each time-step, the relative position between pedestrians is processed through embed layer and LSTM to learn the social relation representation of current time-step.

tion features between pedestrians. The social relation attention is calculated by:

$$\alpha_{ij}^t = \frac{\exp W^{at}\left[r_{ij}^t; h_i^{t-1}; h_j^{t-1}\right]}{\sum_{k\in N(i)} \exp W^{at}\left[r_{ik}^t; h_i^{t-1}; h_k^{t-1}\right]} \tag{3}$$

where the r_{ij}^t is the hidden state of the R-LSTM, which represents the social relation feature between pedestrians i and j. The h_i^{t-1} and h_j^{t-1} are hidden states of pedestrian i and j at time-step $t-1$, which represent the latent motion features of i and j respectively. W^{at} is a weight matrix.

After capturing the social relation attention, we aggregate the latent motion patterns by paying attention to neighboring pedestrians. As illustrated in Fig. 4, we attentively aggregate the hidden states of each pedestrian to obtain social interaction context. The social interaction context of pedestrian i at time-step t is given by:

$$H_i^t = \sum_{j\in N(i)} \alpha_{ij}^t h_j^{t-1} \tag{4}$$

where h_j^{t-1} is the hidden state encoded by history positions of pedestrian j. $N(i)$ is the set of neighbors of pedestrian i, and α_{ij}^t is the attention weight of the pedestrian pair (i,j).

3.4. SRAI-LSTM prediction model

The LSTM-based interactive recurrent structure [18,51,27] has achieved great success in trajectory prediction research. By following these works, we also employ an LSTM as the mainbody to capture the latent motion pattern for each pedestrian, and this LSTM is denoted as motion LSTM (M-LSTM). In our implementation, we use the normalized absolute (Nabs) position [18] which shifts the origin to the latest observed time slot:

$$\begin{aligned} \Delta x_i^t &= x_i^t - x_i^{T_{obs}} \\ \Delta y_i^t &= y_i^t - y_i^{T_{obs}} \end{aligned} \tag{5}$$

The Nabs position is embedded as a fix-length vector by the position embedding layer. The embedding vector and the social interaction context captured from social interaction module are served as input of motion LSTM to capture the current latent motion feature and infer the position of the next time-step. We introduce the following recurrence:

$$e_i^t = \phi\left(\Delta x_i^t, \Delta y_i^t; W_e\right) \tag{6}$$

$$h_i^t = \mathrm{LSTM}\left(h_i^{t-1}, H_i^t, e_i^t; W_l\right) \tag{7}$$

where $\phi(\cdot)$ is an embedding function with ReLU nonlinearity, W_e is the embedding weights. The LSTM weight is denoted by W_l. These parameters are shared among all the pedestrians in the whole scene. H_i^t is the social interaction context tensor which is the output of the social interaction module.

For the deterministic version of SRAI-LSTM, the hidden state h_i^t at time t is used to predict the Nab position $(\Delta\widehat{x}_i^{t+1}, \Delta\widehat{y}_i^{t+1})$ at the next time-step $t+1$directly. For the stochastic version, the hidden state h_i^t and the latent code Z are concatenated and then used for prediction:

$$\begin{aligned} \textit{Deterministic}: &\quad \left[\Delta\widehat{x}_i^{t+1}, \Delta\widehat{y}_i^{t+1}\right]^T = W_p h_i^t \\ \textit{Stochastic}: &\quad \left[\Delta\widehat{x}_i^{t+1}, \Delta\widehat{y}_i^{t+1}\right]^T = W_p\left[h_i^t \oplus Z\right] \end{aligned} \tag{8}$$

where W_p is a weight matrix. $\oplus$is concatenate operation. The latent code $Z$is sampled from standard normal distribution $N(0,1)$. From time $T_{obs}+1$ to T_{pred}, we transform the predicted Nabs positions to absolute positions, which utilized to calculate relative positions to encode social relation.

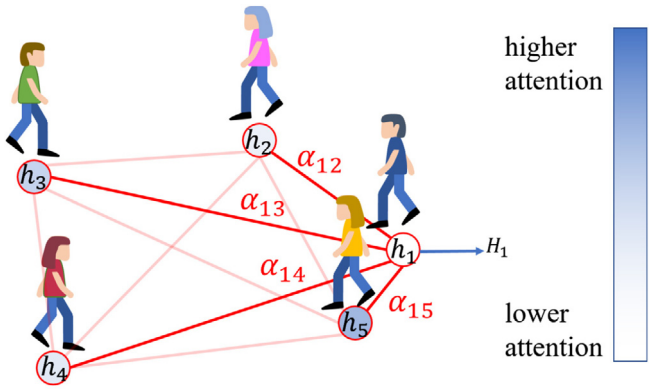


Fig. 4. The red lines between pedestrians represent the existence of human–human interactions. The target pedestrian attentively aggregates the motion features of different neighbors with different attention.

4. Experiments

In this section, we evaluate our method on two public walking pedestrian video datasets: ETH and UCY. These two datasets contain 5 crowd scenes, including ETH, HOTEL, ZARA1, ZARA2, and UNIV. There are 1536 pedestrians and thousands of real-world pedestrian trajectories. All the trajectories are converted to the world coordinate system and then interpolated to obtain values at every 0.4 s.

4.1. Experiment setup

We use the leave-one-out approach similar to that from S-LSTM [11]. Specifically, we train models on four datasets and test on the remaining dataset. We take the coordinates of 8 key frames (3.2s) of the pedestrian as the observed trajectory, and predict the trajectory of the next 12 key frames (4.8s). For each mini-batch, random rotation is employed for data augmentation.

Similar to prior works [12,18], the proposed method is evaluated with two types of metrics as follows:

1. *Average Displacement error(ADE)*: the mean square error(MSE) between the ground-truth trajectory and predicted trajectory over all predicted time steps.
2. *Final Displacement error(FDE)*: the mean square error(MSE) between the ground-truth trajectory and predicted trajectory at the last predicted time steps.

4.2. Implementation details

The parameters of the SRAI-LSTM model are directly learned by minimizing the *L*2loss between the predicted positions and ground truth. All LSTMs in our implementation only have one layer. The dimension of hidden states of all LSTM cells is set to 64. The dimension of embed vector e_{ij}^t in Eq. 1 and e_i^t in Eq. 7 are set to 32. The dimension of latent code *Z*is set to 16. A sliding time window with a length of 20 and a stride size of 1 is adopted to get the training samples. All trajectory segments in the same time window are regarded as a mini-batch, as they are processed in parallel. Adam optimizer is adopted to train models in 300 epochs, with an initial learning rate of 0.001.

4.3. Baseline

We compare the proposed model with the following deterministic models:

1. *Linear*: A linear regressor is used to predict future trajectories by minimizing the least square error.
2. *LSTM*: A LSTM is used for each pedestrian to recurrently predict the future position from historical positions.
3. *S-LSTM* [11]: A trajectory prediction model that combines LSTM with a social pooling layer, which can aggregate hidden states of the neighboring pedestrians.
4. *CIDNN* [52]: A modularized approach for spatio-temporal crowd trajectory prediction with LSTMs.
5. *SR-LSTM* [18]: An interactive recurrent structure with a state refinement module for trajectory prediction.
6. *RSBG* [48]: A seq2seq trajectory prediction approach that uses recursive social behavior graph and GCNs for social interaction modeling.

In addition, we compare the stochastic version of the proposed model with the following multi-modal models.

1. *SGAN* [12]: A multi-modal trajectory prediction model via GAN that uses global pooling to aggregate information from neighboring pedestrians.
2. *Social-Ways* [13]: An improved version of SGAN that attention mechanism used in the pooling module.
3. *STGAT* [10]: A spatio-temporal graph attention network is used to model pedestrians' social interactions for trajectory prediction.
4. *RAMP* [41]: A novel trajectory prediction approach that the forward and backward prediction networks are tightly coupled and satisfying the reciprocal constraint.
5. *Social-STGCNN* [14]: A social spatio-temporal graph convolutional neural network for trajectory prediction.
6. *TPNet* [53]: A unified two-stage motion prediction framework for both vehicles and pedestrians.
7. *NMMP* [42]: A neural motion message passing is proposed to explicitly model the interaction and learn representations for directed interactions between actors.
8. *SILA* [54]: A similarity-based incremental learning algorithm for pedestrian motion prediction.
9. *DSCMP* [43]: A novel future motion predictor, which is able to explicitly model both the spatial and temporal interactions between different agents.

4.4. Quantitative evaluations

4.4.1. Evaluations on ADE/FDE metrics

We compare our method to the state-of-the-art baselines mentioned in Section 4.3. All the stochastic method samples 20 times and reports the best-performed sample. The main results are presented in Table 1,2.

Comparison with deterministic approaches. We compared the deterministic version SRAI-LSTM of the proposed approach with 6 deterministic trajection prediction models. The statistical results illustrate that the SRAI-LSTM model achieves the best performance on the ETH, HOTEL, and ZARA1 sub-datasets while the SR-LSTM [18] model performs best on the rest two sub-datasets. In addition, both SRAI-LSTM and SR-LSTM models have an average ADE of 0.45 on the five datasets. However, our model achieves a lower average FDE (0.93) compared to the SR-LSTM, which is a decrease of 1%.

Comparison with stochastic approaches. Some of the studies specialized in multi-modality of the pedestrian walking process, which could produce plenty of plausible predictions with a single model. We make a simple extension of our model as a stochastic version SRAI-LSTM-S. The proposed model achieves state-of-the-art performance on all sub-datasets. And it achieves the lowest average ADE/ FDE of 0.26/ 0.53 on the five datasets. On ADE metric, compared to the previous state-of-the-art approaches NMMP [42] and DSCMP [43], the SRAI-LSTM-S model improves the performance with 36.59%. And on the FDE metric, the SRAI-LSTM-S model improves the performance with 29.33% compared with the previous state-of-the-art model Social-STGCNN [14].

4.4.2. Model parameter amount and inference speed

To evaluate the inference speed, we list out the size of parameters and inference speed comparisons between our model and publicly available models which we could bench-mark against. Data fragments are densely sampled from the sequential data with the time stride of 1 and the window size of 20 (Tobs(8) +Tpred(12)). For evaluating the inference speed, we treat each fragment as a batch and calculate the average time over all batches. For stochastic models SGAN, STGAT, and NMMP, we only calculate the inference time for one sampling. In particular, the parameter size of the NMMP model is different across the five sub-datasets, the value listed in the table is the average value. In the proposed model, each pair of pedestrians needs to be encoded to capture the feature of social relation. Thus, it costs more time compared with SGAN, STGAT and NMMP models, especially in crowded scenes. Since the SR-LSTM model refines the hidden state of two times at each time step, the inference time of the model is increased. Although the proposed model achieved state-of-the-art performance, its inference speed needed to improve in future work.

Table 1
Comparison with baseline models. SRAI-LSTM-S denotes the stochastic version of SRAI-LSTM. The results of stochastic models are calculated on 20 samples.

Deterministic	Publications	Performance (ADE/FDE) ↓					
		ETH	HOTEL	UNIV	ZARA1	ZARA2	AVERAGE
LR		1.33/2.94	0.39/0.72	0.82/0.59	0.62/1.21	0.77/1.48	0.79/1.59
LSTM		1.13/2.39	0.69/1.47	0.73/1.60	0.64/1.43	0.54/1.21	0.75/1.62
S-LSTM [11]	CVPR'16	1.09/2.35	0.79/1.76	0.67/1.40	0.47/1.00	0.56/1.17	0.72/1.54
CIDNN [52]	CVPR'18	1.25/2.32	1.31/2.36	0.90/1.86	0.50/1.04	0.51/1.07	0.89/1.73
SR-LSTM [18]	CVPR'19	0.63/1.25	0.37/0.74	**0.51/1.10**	0.41/0.90	**0.32/0.70**	**0.45**/0.94
RSBG [48]	CVPR'20	0.80/1.53	0.33/0.64	0.59/1.25	0.40/0.86	0.30/0.65	0.48/0.99
SRAI-LSTM		**0.59/1.16**	**0.29/0.56**	0.55/1.19	**0.37/0.82**	0.43/0.93	**0.45/0.93**
Stochastic		ETH	HOTEL	UNIV	ZARA1	ZARA2	AVERAGE
SGAN [12]	CVPR'18	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-Ways [13]	CVPRW'19	0.39/0.64	0.39/0.66	0.55/1.31	0.44/0.64	0.51/0.92	0.46/0.83
STGAT [10]	ICCV'19	0.65/1.12	0.35/0.66	0.52/1.10	0.34/0.69	0.29/0.60	0.43/0.83
RAMP [41]	CVPR'20	0.69/1.24	0.43/0.87	0.53/1.17	0.28/0.61	0.28/0.59	0.44/0.90
Social-STGCNN [14]	CVPR'20	0.64/1.11	0.49/0.85	0.44/0.79	0.34/0.53	0.30/0.48	0.44/0.75
TPNet [53]	CVPR'20	0.84/1.73	0.24/0.46	0.54/0.94	0.33/0.75	0.26/0.60	0.42/0.90
NMMP [42]	CVPR'20	0.61/1.08	0.33/0.63	0.52/1.11	0.32/0.66	0.29/0.61	0.41/0.82
SILA [54]	CVPRW'20	0.56/1.23	0.27/0.63	0.55/1.25	0.29/0.63	0.32/0.72	0.39/0.89
DSCMP [43]	ECCV'20	0.66/1.21	0.27/0.46	0.50/1.07	0.33/0.68	0.28/0.60	0.41/0.80
SRAI-LSTM-S		**0.32/0.59**	**0.18/0.34**	**0.35/0.72**	**0.24/0.51**	**0.23/0.50**	**0.26/0.53**

Table 2
Comparisions of parameter amount and inference speed on ETH & UCY datasets. All models evaluated on Nvidia GTX2080Ti GPU.

Model	Parameters (k)	Speed (ms/batch)
SGAN [12]	46.4	8.80
SR-LSTM [18]	64.9	22.82
STGAT [10]	44.6	11.05
NMMP [42]	115.8	13.14
Social-STGCNN [14]	6.7	2.27
SRAI-LSTM	67.1	19.00

4.5. Ablation studies

In this section, we will illustrate the effectiveness of each part of the SRAI-LSTM model through additional experiments.

4.5.1. Ablation study of social relation encoder

In the social relation encoder module, we adopt an LSTM model to learn the feature of social relation between pedestrians from their relative positions. To verify the performance of social relation encoder, an ablation study of the key parameters is conducted. Table 3 lists the comparisions of four combinations of the dimension values of embedded vector and the hidden state vector. The statistical results show that the combination (ID: 2) achieves the best performance on ETH, ZARA1, and ZARA2 sub-datasets and achieve the lowest average ADE (0.26) & FDE (0.53). Meanwhile, the combination (ID: 4) performs best on the remaining two sub-datasets and achieves the top2 performance on the average ADE/FDE. Therefore, it can be argued that setting the hidden state dimension to twice the dimension of the embedded vector is more conducive to improving the performance of SRAI-LSTM model.

4.5.2. Ablation study of social interaction module

In this section, we conduct the ablation study of social interaction module to verify the effectiveness of social interaction modeling via social relation attention. As shown in Table 4, we first compare the proposed model (ID: 4) with the none-attention version (ID: 1), and the former performes better in ETH, HOTEL, UNIV, and ZARA1 sub-datasets. The proposed social relation attention is calculated from the feature of social relation between pedestrians and their latent motion features. Therefore, we also compare the attention-based versions (IDs: 2, 3) which are acquired by social relation features and motion features, respectively. The results manifest that the attention captured by coupling the two features is more effective than that captured by either feature in trajectory prediction.

4.5.3. Additional experiments of sampling times

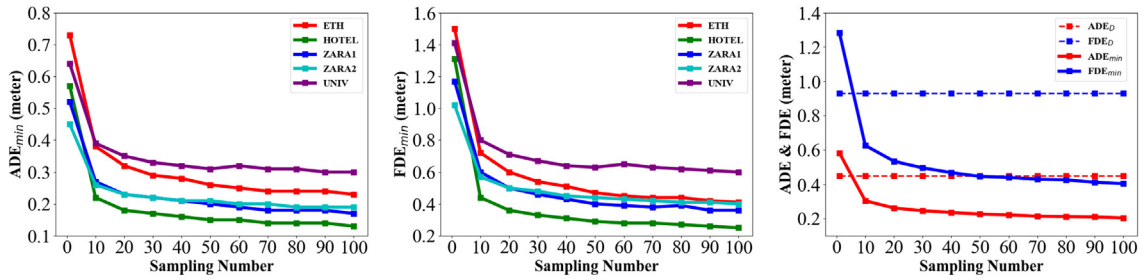
In this section, additional experiments are conducted to investigate the correlation between the effect of trajectory prediction and the number of sampling. For each sampling number, we record the minimum ADE and FDE of each trajectory sample and the average value of all the samples of each sub-dataset are shown in Fig. 5(a) and (b). It is helpful to discover the better predicted trajectory as the sampling times increase. However, when the number of sampling times exceeds 50, the minimum ADE and FDE values tend to be stable, which means that it is difficult to get the better predicted trajectory, even if the number of sampling times is increased. The comparisons of the deterministic and stochastic versions of the proposed model are shown in Fig. 5(c). In the case of sampling once, the performance of the deterministic version is 22.41% and 27.34% better than the stochastic version on ADE & FDE metrics. However, in the case of sampling multiple times, the stochastic version has a better performance than the deterministic version. The execution of sampling multiple times means that the model needs longer inference time.

Table 3
Comparisions on the dimensions combinations of embedding vector and hidden state for the social relation encoder module of Stochastic version SRAI-LSTM-S.

ID	Embedding dim	Hidden dim	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVERAGE
1	32	32	0.37/0.70	0.19/0.40	0.36/0.76	0.26/0.54	0.25/0.54	0.29/0.59
2	32	64	**0.32/0.59**	0.18/0.34	0.35/0.72	**0.24/0.51**	**0.23/0.50**	**0.26/0.53**
3	64	64	0.35/0.70	0.17/0.33	0.37/0.78	0.29/0.63	0.25/0.55	0.29/0.60
4	64	128	0.37/0.71	**0.17/0.32**	**0.33/0.66**	0.29/0.62	0.24/0.52	0.28/0.57

Table 4
Ablation study of the attentions used in social interaction module. Comparisions of attention versions (IDs: 2, 3, 4) with non-attention version (ID: 1). The attentions are calculated by different features. **MF** denotes using motion features, **SRF** denotes using social relation feature.

ID	Feature Components		Performance (ADE/FDE)					
	MF	SRF	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVERAGE
1	–	–	0.69/1.38	0.45/0.99	0.59/1.28	0.48/1.05	**0.36/0.81**	0.51/1.10
2	✓	–	0.65/1.28	0.33/0.63	0.58/1.23	0.44/0.96	0.41/0.88	0.48/1.00
3	–	✓	0.62/1.23	0.34/0.66	0.57/1.22	0.44/0.95	0.40/0.88	0.47/0.99
4	✓	✓	**0.59/1.16**	**0.29/0.56**	**0.55/0.82**	**0.37/0.82**	0.43/0.93	**0.45/0.93**



(a) ADEs on 5 sub-datasets (b) FDEs on 5 sub-datasets (c) Average ADEs & FDEs of 5 sub-datasets

Fig. 5. (a) ADEs on 5 sub-datasets with different numbers of sampling. (b) FDEs on 5 sub-datasets with different numbers of sampling. (c) The average ADEs and FDEs of 5 sub-datasets with different numbers of sampling. The $ADE_D and FDE_D$ represent the prediction result of deterministic version of SRAI-LSTM model.

4.6. Qualitative evaluations

As mentioned before, the quantitative results show that SRAI-LSTM outperforms state-of-art models in terms of ADE/FDE metrics. We now qualitatively analyze how the SRAI-LSTM model successfully captures social interactions. We compare the performances of SGAN, SR-LSTM, and our model in four common social scenarios. And then, we provide comparisons between our model with SR-LSTM in two different social scenarios of group. Besides, the visualization of social relation attention demonstrates the effectiveness of the proposed model on social interaction modeling. Finally, we analyzed the failure cases in two different scenarios.

4.6.1. Individual interaction scenarios

We consider four types of social scenarios where people have to consider possible social interactions to avoid collision (see Fig. 6).

Parallel Walking. (Column 1) It is very common for two individuals to walk in parallel in real-scenarios. For example, the behaviors of classmates going to school together and couples shopping together. There are 3 parallel walking scenarios from ETH, ZARA1, and UNIV sub-datasets shown in Fig. 6(a). For the first scenario, the SGAN model cannot successfully predict future trajectories. The SR-LSTM model and the proposed model can successfully predict the future trajectories in the two scenarios. Compared with the SR-LSTM model, the trajectories predicted by SRAI-LSTM are closest to ground truth. It is worth noting the fourth scenario, the pair of pedestrians changed their original directions of movement in future trajectory. The SGAN model and SR-LSTM model still predict the future trajectories in the original direction of movement. Our model successfully learned the change of pedestrians' movement directions and predicted the future trajectories. Unfortunately, our model does not learn the speed of pedestrian's movement that the predicted position of the final time-step is still a little bit away from ground truth. To a certain extent, our prediction is still successful.

People Merging. (Column 2) In hallways or on roads, it is common for people coming from different directions to merge and walk towards a common destination. People adopt various ways such as walking speed and slowing down to avoid colliding while continuing towards their destinations. As the first scenario shown in Fig. 6(b), the lower pedestrian walk speed up to avoid collision with the upper pedestrian. The prediction of the SGAN model shows that the lower person avoids collision by changing the direction of movement, which is inconsistent with ground truth. The prediction of the SR-LSTM model shows that the lower pedestrian approaches the upper pedestrian and then adjusts the direction of movement to move away, which is also inconsistent with ground truth. We predict speed behavior and successfully predict the future trajectory, which closely matches with the ground truth trajectory. In other two scenarios, the predicted trajectories by SRAI-LSTM also closely match the ground truth unlike the deviation we see in SGAN and SR-LSTM.

Person Following. (Column 3) People tend to follow the people in front when walking toward a common destination. But if the person in front walks slowly, the person at the back has to bypass them to reach the destination. For the first scenario shown in Fig. 6 (c), the three pedestrians have the common destination, but the walking speed of the pair of pedestrians in front is slower than the pedestrian behind. In this situation, the pedestrian behind chose to pass over them from the left side of the pair. For the pedestrian behind, the predicted trajectories of these three models have all been detoured, but only our prediction is close to the ground truth. Although the predicted trajectory does not reach the specified position, the predicted movement direction is correct. In other two scenarios, the trajectories predicted by our model are also the closest to the ground truth.

People Meeting. (Column 4) On the way to the destination, there will be other people coming from the opposite direction. For the face-to-face meeting, one has to adjust the direction of movement temporarily to avoid collision. In other cases, it is not necessary. As the first scenario shown in Fig. 6(d), without avoidance, the two pedestrians would pass by. The upper pedestrian's trajectories predicted by SGAN and SR-LSTM changed the direction of movement to avoid collision, which deviated from the ground truth. The trajectory predicted by our model of this person is closest to the ground truth. But the predicted trajectories for the lower pedestrian are all shorter than ground truth. For other scenarios, our model can predict the trajectory closest to ground truth.

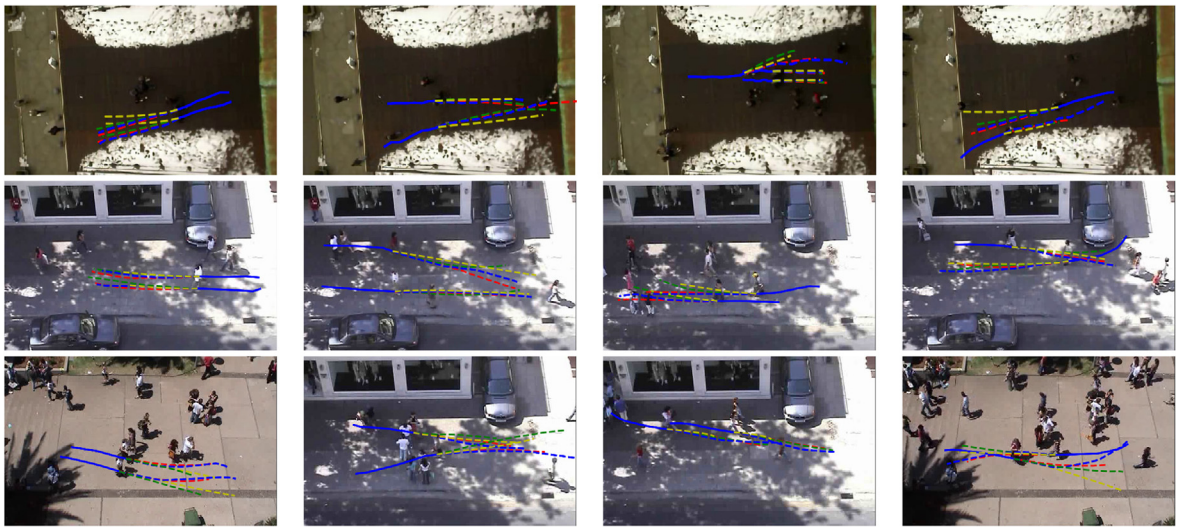


Fig. 6. Comparisons of our method with SGAN and SR-LSTM in 4 common individual interaction scenarios. The examples in each column belong to the same type of social scenario, from left to right: parallel walking, people merging, person following, and people meeting. For each case, the blue solid lines represent the observed trajectories, the dashed lines are the future trajectories (blue: ground truth, yellow: SGAN, green: SR-LSTM, red: our model). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.6.2. Group interaction scenarios

When there are many pedestrians, they unconsciously form a group, especially acquaintances. We compare our model with the SR-LSTM model in two common group scenarios. The scenarios include group parallel walking and group meeting.

Group parallel walking. The group parallel walking behaviors such as several classmates going to school together or several friends shopping together. The group walking scenarios from ETH and ZARA1 sub-datasets are shown in Fig. 7(a) and (b). Row 1 shows the predictions of the SR-LSTM model, and row 2 shows the predictions of our model. For the left scenario, the trajectories predicted by SR-LSTM show that pedestrians 3 and 4 are gradually getting closer together, which means that the two collide. Although there is a deviation between the trajectory predicted by our model and the ground truth, there is no collision. For the scenario in column 2, there are five pedestrians walking together. There are various deviations between the trajectories predicted by SR-LSTM with the ground truth. The same goes for our predicted trajectories. Although our predicted trajectories have various deviations, they are closer to ground truth. It also shows that our predicted pedestrian's walking speed is faster than in ground truth.

Group meeting. There are two scenarios of group meeting cases shown in Fig. 7(c) and (d). In these cases, two pairs of pedestrians met from opposite directions. For the scenario in column 3, pedestrians 1 and 2's trajectories predicted by our model are closer to ground truth than SR-LSTM. Pedestrian 3 changed the direction of movement to avoid collision with pedestrian 2, and pedestrian 4 as a partner of pedestrian 3 also changed. The pedestrian 3's trajectory predicted by our model is closer to the ground truth. The trajectories of pedestrian 4 predicted by the two models are quite different from the ground truth. For the scenario in column 4, the pair of pedestrians 1 and 2 change the direction of their movement to avoid collision with the other pair, and then return to the original direction. Our prediction only changed the direction of pedestrians 1 and 2, but did not return to the original direction, which is the reason for the deviation of the prediction trajectory. However, the predicted trajectory of pedestrian 3 matches with ground truth very well. The trajectory of pedestrian 4 predicted by SR-LSTM is better than our prediction.

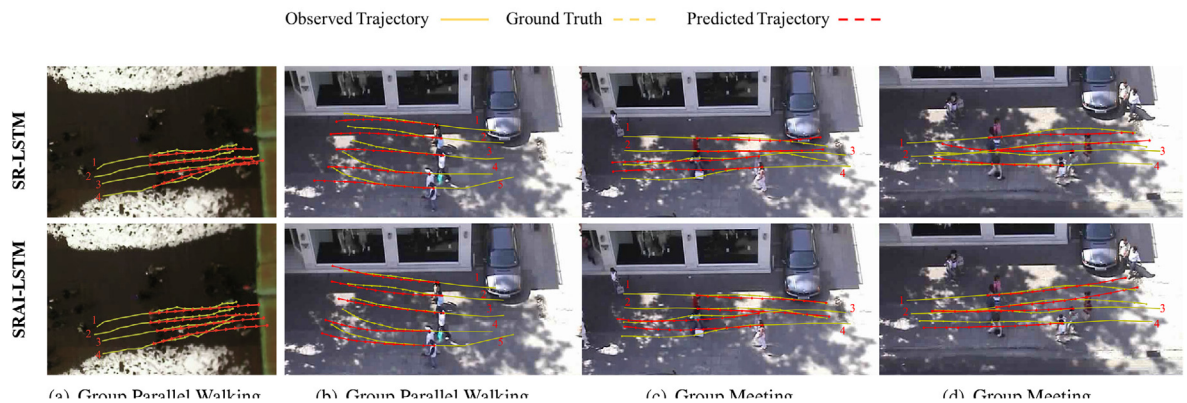


Fig. 7. Comparisons of our method with SR-LSTM in 4 group interaction scenarios. (a) and (b) show the predictions in group parallel walking scenarios. (c) and (d) show the predictions in group meeting scenarios. The predictions of SR-LSTM and SRAI-LSTM are shown in row 1 and row 2 respectively.

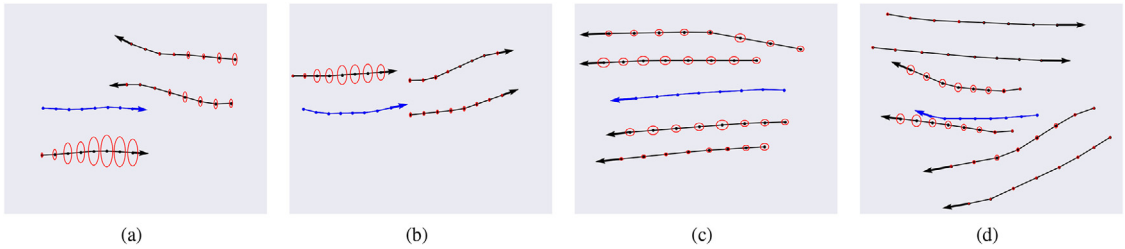


Fig. 8. Visualizations of the social relation attention weights of 4 group interaction scenarios. For each scene, the trajectory of the target pedestrian is represented by a blue line, and its neighboring pedestrians' trajectories are represented by black lines. The circles on trajectories represent the attention weights and the radius of circle proportional to attention weight. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.6.3. Attention weights visualization

To verify the reasonability of our proposed social relation attention, we visualize the learned attention weights of social interaction module. As shown in Fig. 8, we choose four different types of group interaction scenarios to visualize the social relation attention weights. And for a better demonstration, we only show the attention weights from one interested target person to its neighborhood pedestrians. For each scene, the trajectoris of the target pedestrian and its neighbors are represented by blue and black lines, respectively. The circles on trajectories represent the attention weights and the radius of circle proportional to attention weight. As shown in Fig. 8(a)–(c), when the target pedestrian is walking with his/her companion, our model can successfully learn to pay higher attention to the companion. However, as shown in Fig. 8(d), the target pedestrian is walking alone, he pays higher attention to the pedestrian who has the same motion pattern as others. The attention in our model is calculated by the social relation feature which is acquired from the relative positions of pedestrians. Thus, pedestrians with the same motion pattern can easily be mistaken for companions walking in a group.

4.6.4. Failure cases scenarios

The above two subsections show the successful cases of the model in individual interaction scenarios and group interaction scenarios. However, our model cannot successfully predict the future trajectory of pedestrians in some special scenarios. Two failure case scenarios are shown in Fig. 9. The target pedestrian in Fig. 9(a) walks towards the wall and chooses to stop by the wall. Our model cannot understand this motivation from the pedestrian's observed trajectory, and thus incorrectly predicts the future trajectory. The reason for the failure case shown in Fig. 9(b) is also that the model cannot understand the pedestrian's motivation from the observed trajectory, which leads to the failure of the prediction. If the model can understand the scene structure information, the above failure cases may not exist. However, the thorough solution to the problem lies in accurately understanding the motion motivation of pedestrians, which is also a difficulty in trajectory prediction research. And it is also the goal of our future work.

Fig. 9. Scenarios of failure case. The blue solid line represents the observed trajectory, the yellow dashed line represents groundtruth, and the red dashed line represents the predicted trajectory. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5. Conclusions and future works

In this paper, we propose a novel interactive recurrent structure SRAI-LSTM for predicting the future trajectories of pedestrians in the crowd. Quantitative evaluations on public datasets prove that the proposed model is superior to state-of-art models. The qualitative results of some individual interaction scenarios and group interaction scenarios demonstrate the effectiveness of our method in social interaction modeling.

The proposed SRAI-LSTM model successfully predicts the future trajectories of pedestrians in most scenarios. As shown in Section 4.6.4, the proposed model has failed to successfully predict in some scenarios. The reason is that the SRAI-LSTM model only models social interaction among pedestrians but ignores the scene layout information. In the future, we will focus on modeling human-scene interactions and expect to improve the prediction performance by coupling human-scene interaction with our SRAI-LSTM prediction model. Besides, how to effectively perceive the intentions of pedestrians' movement is also a difficult point that we need to solve.

CRediT authorship contribution statement

Yusheng Peng: Conceptualization, Methodology, Investigation, Writing - original draft, Writing - review & editing. **Gaofeng Zhang:** Conceptualization, Methodology, Writing - review & editing, Validation, Funding acquisition. **Jun Shi:** Methodology, Writing - review & editing. **Benzhu Xu:** Conceptualization, Supervision. **Liping Zheng:** Conceptualization, Supervision, Validation, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (61972128), the Fundamental Research Funds for the Central Universities of China (Grant No. PA2019GDPK0071).

References

[1] S. Zhang, D. Cheng, Y. Gong, D. Shi, X. Qiu, Y. Xia, Y. Zhang, Pedestrian search in surveillance videos by learning discriminative deep features, Neurocomputing 283 (2018) 120–128, https://doi.org/10.1016/j.neucom.2017.12.042.

[2] X. Zhang, X. Yang, W. Zhang, G. Li, H. Yu, Crowd emotion evaluation based on fuzzy inference of arousal and valence, Neurocomputing 445 (2021) 194–205, https://doi.org/10.1016/j.neucom.2021.02.047.

[3] K. Saleh, M. Hossny, S. Nahavandi, Spatio-temporal densenet for real-time intent prediction of pedestrians in urban traffic environments, Neurocomputing 386 (2020) 317–324, https://doi.org/10.1016/j.neucom.2019.12.091.

[4] Y. Luo, P. Cai, A. Bera, D. Hsu, W.S. Lee, D. Manocha, Porca: Modeling and planning for autonomous driving among many pedestrians, IEEE Robot. Autom. Lett. 3 (2018) 3418–3425, https://doi.org/10.1109/LRA.2018.2852793.

[5] T. Obo, Y. Nakamura, Intelligent robot navigation based on human emotional model in human-aware environment, in: 2019 International Conference on Machine Learning and Cybernetics (ICMLC), 2019, pp. 1–6, https://doi.org/10.1109/ICMLC48188.2019.8949247.

[6] C. Mavrogiannis, A.M. Hutchinson, J. Macdonald, P. Alves-Oliveira, R.A. Knepper, Effects of distinct robot navigation strategies on human behavior in a crowded environment, in: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, 2019, pp. 421–430.

[7] D. Helbing, P. Molnar, Social force model for pedestrian dynamics, Phys. Rev. E 51 (1998) 4282–4286, https://doi.org/10.1103/PhysRevE.51.4282.

[8] S. Yi, H. Li, X. Wang, Pedestrian behavior modeling from stationary crowds with applications to intelligent surveillance, IEEE Trans. Image Process. 25 (2016) 4354–4368, https://doi.org/10.1109/TIP.2016.2590322.

[9] S. Yi, H. Li, X. Wang, Understanding pedestrian behaviors from stationary crowd groups, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3488–3496, https://doi.org/10.1109/CVPR.2015.7298971.

[10] Y. Huang, H. Bi, Z. Li, T. Mao, Z. Wang, Stgat: modeling spatial-temporal interactions for human trajectory prediction, in: in: 2019 IEEE International Conference on Computer Vision (ICCV), IEEE, 2019, pp. 6271–6280.

[11] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, S. Savarese, Social lstm: human trajectory prediction in crowded spaces, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 961–971. doi: 10.1109/CVPR.2016.110.

[12] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, A. Alahi, Social gan: socially acceptable trajectories with generative adversarial networks, in: in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2018, pp. 2255–2264.

[13] J. Amirian, J.B. Hayer, J. Pettre, Social ways: learning multi-modal distributions of pedestrian trajectories with gans, in: in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2019, pp. 2964–2972.

[14] A. Mohamed, K. Qian, M. Elhoseiny, C. Claudel, Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020, pp. 14412–14420.

[15] H. Zhou, D. Ren, H. Xia, M. Fan, X. Yang, H. Huang, Ast-gnn: An attention-based spatio-temporal graph neural network for interaction-aware pedestrian trajectory prediction, Neurocomputing 445 (2021) 298–308, https://doi.org/10.1016/j.neucom.2021.03.024.

[16] T. Fernando, S. Denman, S. Sridharan, C. Fookes, Soft+ hardwired attention: an lstm framework for human trajectory prediction and abnormal event detection, Neural Netw. (2018) 466–478, https://doi.org/10.1016/j.neunet.2018.09.002.

[17] A. Vemula, K. Muelling, Social attention: modeling attention in human crowds, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 1–7.

[18] P. Zhang, W. Ouyang, P. Zhang, N. Zheng, Sr-lstm: state refinement for lstm towards pedestrian trajectory prediction, in: in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 12077–12086.

[19] N. Kamra, H. Zhu, D. Trivedi, M. Zhang, Y. Liu, Multi-agent trajectory prediction with fuzzy query attention, in: in: 2020 34th International Conference on Neural Information Processing Systems (NeurIPS), NIPS, 2020, pp. 1–16, URL: https://proceedings.neurips.cc/paper/2020/hash/fe87435d12ef7642af67d9bc82a8b3cd-Abstract.html.

[20] C. He, B. Yang, L. Chen, G. Yan, An adversarial learned trajectory predictor with knowledge-rich latent variables, in: in: Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Springer, 2020, pp. 42–53.

[21] A. Lerner, Y. Chrysanthou, D. Lischinski, Crowds by example, Comput. Graph. Forum (2007) 655–664, https://doi.org/10.1111/j.1467-8659.2007.01089.x.

[22] S. Pellegrini, A. Ess, K. Schindler, V.G.L., You'll never walk alone: modeling social behavior for multi-target tracking, in: 2009 IEEE International Conference on Computer Vision (ICCV), 2009, pp. 261–268. doi: 10.1109/ICCV.2009.5459260.

[23] E.A. John, Forecasting, structural time series and the kalman filter, Technometrics 34 (1992) 496–497, https://doi.org/10.1080/00401706.1992.10484972.

[24] D. Ellis, E. Sommerlade, I. Reid, Modelling pedestrian trajectory patterns with gaussian processes, in: in: 2009 IEEE International Conference on Computer Vision Workshops (ICCV), IEEE, 2009, pp. 1229–1234.

[25] K.M. Kitani, B.D. Ziebart, J.A. Bagnell, M. Hebert, Activity forecasting, in: in: 2012 European Conference on Computer Vision (ECCV), Springer, 2012, pp. 201–214.

[26] I. Hasan, F. Setti, T. Tsesmelis, D.B. Alessio, F. Galasso, M. Cristani, Mx-lstm: Mixing tracklets and vislets to jointly forecast trajectories and head poses, in: in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2018, pp. 6067–6076.

[27] Y. Xu, J. Yang, S. Du, Cf-lstm: Cascaded feature-based long short-term networks for predicting pedestrian trajectory, in: The 34th AAAI Conference on Artificial Intelligence (AAAI), vol. 34, 2020, pp. 12541–12548. URL: 10.1609/aaai.v34i07.6943.

[28] T. Fernando, S. Denman, S. Sridharan, C. Fookes, Gd-gan: Generative adversarial networks for trajectory prediction and group detection in crowds, in: 2018 Asian Conference on Computer Vision, vol. 11361, 2019, pp. 314–330. doi: 10.1007/978-3-030-20887-5.

[29] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, S. Savarese, Sophie: An attentive gan for predicting paths pompliant to social and physical constraints, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 1349–1358.

[30] P. Dendorfer, A. Ošep, L. Leal-Taixé, Goal-gan: Multimodal trajectory prediction based on goal position estimation, in: 2018 Asian Conference on Computer Vision, vol. 13623, 2021, pp. 405–420. doi: 10.1007/978-3-030-69532-3_25.

[31] K. Mangalam, H. Girase, S. Agarwal, K. Lee, E. Adeli, J. Malik, A. Gaidon, It is not the journey but the destination: Endpoint conditioned trajectory prediction, in: 2020 Eupean Conference on Computer Vision (ECCV), vol. 1, 2020, pp. 759–776. doi: 10.1007/978-3-030-58536-5.

[32] Y. Yao, M. Atkins, E.and Johnson-roberson, R. Vasudevan, X. Du, Bitrap: Bi-directional pedestrian trajectory prediction with multi-modal goal estimation, IEEE Robot. Autom. Lett. 6 (2021) 1463–1470. doi: 10.1109/LRA.2021.3056339.

[33] D. Xiong, Spatial-temporal block and lstm network for pedestrian trajectories prediction, arXiv:2009.10468 (2020). URL: https://arxiv.org/ftp/arxiv/papers/2009/2009.10468.pdf.

[34] C. Wang, S. Cai, G. Tan, Graphtcn: Spatio-temporal interaction modeling for human trajectory prediction, 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3449-3458, doi: 10.1109/WACV48630.2021.00349.

[35] C. Yu, X. Ma, J. Ren, H. Zhao, S. Yi, Spatio-temporal graph transformer networks for pedestrian trajectory prediction, in: 2020 European Conference on Computer Vision (ECCV), volume 2, 2020, pp. 507–523. doi: 10.1007/978-3-030-58610-2.

[36] Y. Ye, X. Weng, Y. Ou, K. Kitani, AgentFormer: Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting, 2021 IEEE International Conference on Computer Vision (ICCV), 2021, pp. 9813-9823.

[37] J. Tang, X. Shu, R. Yan, L. Zhang, Coherence constrained graph lstm for group activity recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2019) 1–12, https://doi.org/10.1109/TPAMI.2019.2928540.

[38] X. Shu, L. Zhang, Y. Sun, J. Tang, Host-parasite: Graph lstm-in-lstm for group activity recognition, IEEE Trans. Neural Netw. Learn. Syst. 32 (2021) 663–674, https://doi.org/10.1109/TNNLS.2020.2978942.

[39] X. Shu, L. Zhang, G.-J. Qi, W. Liu, J. Tang, Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction, IEEE Trans. Pattern Anal. Mach. Intell. (2021) 1–16, https://doi.org/10.1109/TPAMI.2021.3050918.

[40] A. Alahi, V. Ramanathan, F.-F. Li, Socially-aware large-scale crowd forecasting, in: in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2014, pp. 2211–2218.

[41] H. Sun, Z. Zhao, Z. He, Reciprocal learning networks for human trajectory prediction, in: in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020, pp. 7414–7423.

[42] Y. Hu, B. Chen, Y. Zhang, X. Gu, Collaborative motion prediction via neural motion message passing, in: in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020, pp. 6318–6327.

[43] C. Tao, Q. Jiang, L. Duan, P. Luo, Dynamic and static context-aware lstm for multi-agent motion prediction, in: 2020 Eupean Conference on Computer Vision (ECCV), vol. 12366, 2020, pp. 547–563. doi: 10.1007/978-3-030-58589-1_33.

[44] F. Li, Q. Li, Z. Li, Z. Huang, X. Chang, J. Xia, A personal location prediction method based on individual trajectory and group trajectory, IEEE Access 7 (2019) 92850–92860, https://doi.org/10.1109/ACCESS.2019.2927888.

[45] W. Liang, W. Zhang, Learning social relations and spatiotemporal trajectories for next check-in inference, IEEE Trans. Neural Netw. Learn. Syst. (2020) 1–11, https://doi.org/10.1109/TNNLS.2020.3016737.

[46] F. Li, Z. Gui, Z. Zhang, D. Peng, S. Tian, K. Yuan, Z. Sun, H. Wu, J. Gong, Y. Lei, A hierarchical temporal attention-based lstm encoder-decoder model for individual mobility prediction, Neurocomputing 403 (2020) 153–166, https://doi.org/10.1016/j.neucom.2020.03.080.

[47] M. Zhang, X. Liu, W. Liu, A. Zhou, H. Ma, T. Mei, Multi-granularity reasoning for social relation recognition from images, in: in: 2019 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2019, pp. 1618–1623.

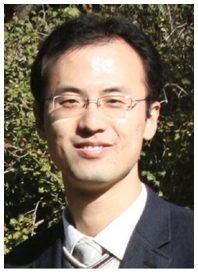
[48] J. Sun, Q. Jiang, C. Lu, Recursive social behavior graph for trajectory prediction, in: in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020, pp. 657–666.

[49] N. Bisagno, B. Zhang, N. Conci, Group LSTM: group trajectory prediction in crowded scenarios, in: in: 2018 European Conference on Computer Vision (ECCV), vol. 11131, Springer Verlag, 2018, pp. 213–225.

[50] N. Bisagno, C. Saltori, B. Zhang, F.G.B.D. Natale, N. Conci, Embedding group and obstacle information in lstm networks for human trajectory prediction in crowded scenes, Comput. Vis. Image Underst. 203 (2021) 1–9, https://doi.org/10.1016/j.cviu.2020.103126.

[51] P. Zhang, J. Xue, P. Zhang, N. Zheng, Social-aware pedestrian trajectory prediction via states refinement lstm, IEEE Trans. Pattern Anal. Mach. Intell. (2020) 1–18, https://doi.org/10.1109/TPAMI.2020.3038217.
[52] Y. Xu, Z. Piao, S. Gao, Encoding crowd interaction with deep neural network for pedestrian trajectory prediction, in: in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2018, pp. 5275–5284.
[53] L. Fang, Q. Jiang, J. Shi, B. Zhou, Tpnet: Trajectory proposal network for motion prediction, in: in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020, pp. 6796–6805.
[54] G. Habibi, N. Jaipuria, J.P. How, Sila: An incremental learning approach for pedestrian trajectory prediction, in: in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR), IEEE, 2020, pp. 4411–4421.

Yusheng Peng received the M.S. degree in Computational Mathematics from Anqing Normal University, China, in 2017. He is currently pursuing an Ph.D. degree at Hefei University of Technology, China. His current research interests include trajectory prediction, crowd evacuation and simulation, and software engineering.



Gaofeng Zhang received the Ph.D. degree in ICT from Swinburne University of Technology (SUT), Australia, in 2013. Now, he is an associate professor at Hefei University of Technology, China. His research interests include cloud/edge computing, software security, public safety and software engineering.

Jun Shi received the Ph.D. degree in pattern recognition and intelligent systems from Beijing University of Aeronautics and Astronautics, China, in 2011. Now, he is a lecturer at Hefei University of Technology, China. His research interests include machine learning, medical image analysis and remote sensing image understanding.

Benzhu Xu received the master's and Ph.D. degrees in computer science from Hefei University of Technology, China. He is currently an associate Professor with the School of Software, Hefei University of Technology. His research interests include Software Engineering and Computer-Aided Design.

Liping Zheng received the master's and Ph.D. degrees in computer science from Hefei University of Technology, China. He is currently a professor with the School of Software, Hefei University of Technology. His research interests include crowd simulation and computer graphics.