SRGAT: Social Relational Graph Attention Network for Human Trajectory Prediction

Yusheng Peng[1], Gaofeng Zhang[2,3], Xiangyu Li[1], and Liping Zheng[1,2,3(✉)]

[1] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China
wisionpeng@mail.hfut.edu.cn, zhenglp@hfut.edu.cn
[2] School of Software, Hefei University of Technology, Hefei 230601, China
[3] Anhui Province Key Laboratory of Industry Safety and Emergency Technology, Hefei 230601, China

Abstract. Human trajectory prediction is a popular research of computer vision and widely used in robot navigation systems and automatic driving systems. The existing work is more about modeling the interactions among pedestrians from the perspective of spatial relations. The social relation between pedestrians is another important factor that affects interactions but has been neglected. Motivated by this idea, we propose a Social Relational Graph Attention Network (SRGAT) via seq2seq architecture for human trajectory prediction. Specifically, relational graph attention network is utilized to model social interactions among pedestrians with different social relations and we use a LSTM model to capture the temporal feature among these interactions. Experimental results on two public datasets (ETH and UCY) prove that SRGAT achieves superior performance compared with recent methods and the predicted trajectories are more socially plausible.

Keywords: Social relations · Social Relational Graph Attention Networks (SRGAT) · Social interactions · Trajectory prediction

1 Introduction

As a key technology in robot navigation system and autonomous driving system, the human trajectory prediction has attracted considerable interests from both academia and industry over the past few years. The human trajectory prediction is full of challenges due to the subtle and intricate interactions among pedestrians.

Many scholars have worked to model these subtle and intricate interactions. The earlier works [8,12] attempt to use handcrafted energy functions to model these social interactions in crowded spaces. However, it is still full of challenges to overall consider various social behaviors. With the rapid development of artificial intelligence technology, the deep learning based human trajectory prediction

T. Mantoro et al. (Eds.): ICONIP 2021, LNCS 13109, pp. 632–644, 2021.
https://doi.org/10.1007/978-3-030-92270-2_54

approaches [1,6,10] has achieved great success. In related works, pooling mechanism [1,2,6], attention mechanism [5,16] and graph neural network [9,11] are widely used to model social interactions among pedestrians. Most of these methods model social interaction from the perspective of spatial relations, while the social relations between pedestrians have been neglected.

In view of the limitations of the above methods, we introduce the interpersonal distance to represent social relation. American anthropologist Edward Hall divides interpersonal distance into four kinds: intimate, personal, social and public [7]. We construct an social relational graph among pedestrians according to these four kinds of interpersonal distance. Besides, we introduce relational graph attention network (RGAT) [4] to model the social interactions among pedestrians within different interpersonal distances respectively.

Contributions: We propose a novel Social Relational Graph Attention Network (called SRGAT) with encoder-decoder architecture for trajectory prediction which respectively considers the social influence of neighboring pedestrians within different interpersonal distances. Firstly, we utilize RGAT to model social interactions among pedestrians with different social relations, and then, we adopt a gated attention mechanism to aggregate these social features to acquire social features. Secondly, we use a LSTM to explicitly capture the temporal correlations of these social features. This paper is the first attempt to use RGAT for social interaction modeling in human trajectory prediction. Experimental results demonstrate that the proposed SRGAT model successfully predicts future trajectories of pedestrians.

2 Related Works

2.1 Social Interaction Modeling

Handcrafted rules and energy parameters [8,12] have been used to capture social interactions but fail to generalize properly. In some recent approaches [1,2,6], pooling mechanisms have been used to model social interactions among pedestrians in local or global neighborhoods. In the view that pedestrians have different impacts on social interactions, Fernando et al. [5] introduced an attention mechanism in social interaction modeling. After that, existing approaches [15,16] adopt diverse attention mechanisms to improve performance of trajectory prediction. With the development of graph neural network, graph-based social interaction modeling has been utilized in various pedestrian trajectory models [9,11,18]. In our model, we utilize RGAT to capture spatial interaction features on the social graphs, and the spatial interaction features are fed to an LSTM to model the temporal correlation to capture spatio-temporal interaction features.

2.2 Social Relation Modeling

Social relations are the general term of mutual relations formed by people in the process of common material and spiritual activities, that is, all the relations between people. Recognizing the social relations between people can enable

agents to better understand human behavior or emotions [17]. Sun et al. [14] directly annotate social relations as 0/1 which represents whether pedestrians are in the same group or not. The SRA-LSTM model [13] learned the social relation representations from the relative positions among pedestrians through social relationship encoder. However, the learned representation of social relation is lack of interpretability. In our work, we take the interpersonal distance theory [7] of sociological psychology as a standard and divide social relations according to four kinds of interpersonal distance (intimate, personal, social and public). We model the social interactions of pedestrians of each type of social relationship separately, and integrate the social interactions of four social relationships through gated attention mechanisms.

2.3 Graph Neural Networks

Graph Neural Networks (GNNs) are powerful neural network architecture for machine learning on graphs. Graph neural networks (GNNs) are effective neural networks for processing graph structure data. Recently, the variants of GNN including Graph Convolutional Network (GCN) and Graph Attention Networks (GAT) demonstrate breakthroughs on various tasks like social network prediction, traffic prediction, recommender systems and molecular fingerprints prediction [19]. In the pedestrian trajectory prediction task, GCN is used as message passing to aggregate motion information from nearby pedestrians to model social interactions [3,11,18]. While neighbor pedestrians have different impacts on the target pedestrian, GAT is more suitable for modeling such social interactions and achieved success in predicting future trajectory [9]. Inspired by Relational Graph Attention Network (RGAT) [4], we introduce a novel RGAT to model the social interactions among pedestrians with different social relations, and aggregate the social interaction features of different social relations through a gated attention mechanism.

3 Approach

The goal of human trajectory prediction is to predict the future trajectories from the given past trajectories of pedestrians. The goal of human trajectory prediction is to predict the future trajectories from the given past trajectories of pedestrians. The novel Social Relational Graph Attention Network (SRGAT) via seq2seq structure is proposed in this section (as shown in Fig. 1). For each pedestrian, an LSTM is employed to encode trajectory from relative positions to capture motion feature. Meanwhile, we create social relationship graph among neighbor pedestrians from absolute positions, and the RGAT was utilized to acquire the social interaction feature. Then an extra LSTM is used to encode social interactions of all time steps to capture the spatio-temporal social interaction features. Finally, we employ an LSTM as decoder to predict future positions from encoder feature.

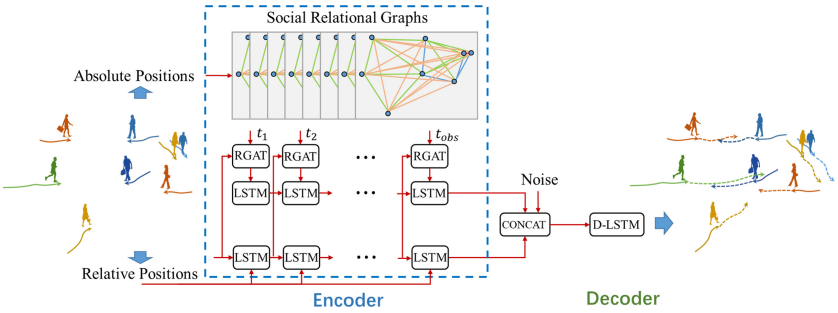

Fig. 1. Illustration of the overall approach. At each time-step, the pedestrians positions are used to calculate relative positions of each other, and the relative positions are processed through embed layer and LSTM to encode the social relationships of each pair of pedestrians. The social relationship attention module models the social interactions by attentively integrating the hidden states of neighbors. Then the social interaction tensor and the embedding vector of each pedestrian's position are treated as inputs of LSTM to output the current hidden states and infer the positions of next time-step.

3.1 Problem Formulation

This paper focuses on addressing the human trajectory prediction in surveillance video crowd scenarios. For better modeling the social interactions among pedestrians, we focus on two-dimensional coordinations of pedestrians in the world coordinate system at specific key frames. For each sample, we assumed that the surveillance video scene involved N pedestrians. Given certain observed positions $\{p_i^t|(x_i^t, y_i^t), t = 1, 2, ..., T_{obs}\}$ of pedestrians i of T_{obs} key frames, our goal is predicting the positions $\{p_i^{t'}|(\widehat{x}_i^{t'}, \widehat{y}_i^{t'}), t' = T_{obs} + 1, T_{obs} + 2, ..., T_{pred}\}$ of future T_{pred} key frames.

3.2 Trajectory Encoding

LSTM is often used to capture latent motion states in pedestrian trajectory prediction models [1,6,16]. By following these works, to capture the unique motion pattern, we also employ an LSTM denoted as Motion Encoder (ME-LSTM) to capture the latent motion pattern for each pedestrian. For each time-step, we embed the relative position into a fixed-length vector e_i^t, and the embedding vector is fed to the LSTM cell as follows:

$$e_i^t = \phi(\Delta x_i^t, \Delta y_i^t; W_e) \tag{1}$$

$$m_i^t = \text{ME-LSTM}(h_i^{t-1}, e_i^t; W_m) \tag{2}$$

where $(\Delta x_i^t, \Delta y_i^t)$ is the relative coordinate of pedestrian i at time-step t to the previous time-step, $\phi(\cdot)$ is an embedding function with ReLU nonlinearity, W_e is the weights of embedding function. The ME-LSTM weight is denoted by W_m. All these parameters are shared by all pedestrians involved in the current scene.

3.3 Social Interaction Modeling

To model the social interactions among pedestrians, pooling mechanism [1,2,6] is used to aggregate hidden states among pedestrians on occupancy map. Besides, GNNs are used to capture social interaction features in recent approaches [3,9] and achieve great successful performance. Inspired by the existed works [3,9], all pedestrians in the scene are treated as nodes on the graph. As illustrated in Fig. 2, the edge between each nodes represents latent social interaction between pedestrians.

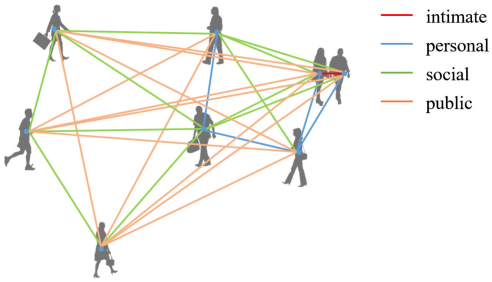

Fig. 2. Pedestrians in the scene are represented by nodes in the graph, and the social interaction between pedestrians is represented by edge between nodes. The different colored lines represent the interaction of different social relations between pedestrians. The intimate, personal, social and public relation are represented by red, blue, green and orange lines, respectively. (Color figure online)

Instead of the above works, we define a social relational graph to represent the social interactions among pedestrians in the scene. As shown in Fig. 2, We divided the social relations between pedestrians into four types: intimate, personal, social, and public [7]. For each social relational graph, all nodes under social relation r *(intimate, personal, social, and public)* represented by an adjacency matrix. The adjacency matrixs are calculated by 4 interpersonal distance ranges: $ranges = [0, 0.45], (0.45, 1.2], (1.2, 3.6], (3.6, 7.5]$:

$$A_{i,j}^{(r)} = \begin{cases} 1, & dist(i,j) \in ranges^{(r)} \\ 0, & otherwise \end{cases} \tag{3}$$

And then, we utilize an relational GAT [4] to model social interactions through the defined social relational graph. The input of RGAT is a graph with 4 relation types and N nodes. For each time-step t, the i^{th} node is represented by the motion feature vector $m_i \in R^F$, and the features of all nodes are summarised in the feature matrix $M^t = [m_1^t m_2^t ... m_N^t] \in R^{N \times F}$. Through the operation of RGAT, we obtain the social interaction features under 4 kinds of relations:

$$\widehat{M^t} = \mathrm{RGAT}(M^t, A^t) \tag{4}$$

where A^t is a summarised Adjacency matrix $A^t = [A^{r_1,t}A^{r_2,t}A^{r_3,t}A^{r_4,t}] \in R^{4\times N\times N}$. The output $\widehat{M}^t$ is a summarised feature matrix $\widehat{M}^t = [\widehat{m}^{r_1,t}\widehat{m}^{r_2,t}\widehat{m}^{r_3,t}\widehat{m}^{r_4,t}] \in R^{4\times N\times F}$.

For each relation$r$, we get a gate value $g^r = 0/1$ to represent whether there have neighbors of pedestrian i under this relation. And then we aggregate 4 social relational interaction features by the gate mechanism:

$$\mathcal{M}_i^t = \sum_{r\in\Re} g^r \cdot \widehat{m}_i^{r,t} \tag{5}$$

We only model the spatial interactions among pedestrians at the time-steps of the observation stage. To learn the temporal correlations between spatial interactions of pedestrians, we propose to employ an extra LSTM to encode these spatial interactions. We term this Social Encoder as SE-LSTM:

$$s_i^t = \text{SE-LSTM}(s_i^{t-1}, \mathcal{M}_i^t; W_s) \tag{6}$$

where W_s is the weight of SE-LSTM which is shared by all pedestrians in this scene.

3.4 Fusion and Prediction

In Encoder component of our proposed model, the ME-LSTM encoder is design to learn motion feature from observated trajectory, and the SE-LSTM encoder is designed to learn spatial-temporal interaction features. These two parts are combined together later to fusion of motion and social interaction features. At the last observated time-step T_{obs}, the encoder features of each pedestrian are represented by two hidden variables ($m_i^{T_{obs}}$, $s_i^{T_{obs}}$) from two LSTMs. The two variables will be fused to served as input of decoder to predict future trajectory. However, because of the uncertainty of pedestrian movement, it is necessary to predict multiple reasonable and socially acceptable trajectories. Thus, the latent code z from $\mathcal{N}(0,1)$ (the standard normal distribution) is added to encoding features. We concatenate three variables in our implementation:

$$d_i^{T_{obs}} = m_i^{T_{obs}} \parallel s_i^{T_{obs}} \parallel z \tag{7}$$

the concatenated state vector $d_i^{T_{obs}}$ then employed as the initial of the decoder LSTM hidden state (termed as D-LSTM). The inferred relative coordinate is given by:

$$d_i^{T_{obs}+1} = \text{D-LSTM}(d_i^{T_{obs}}, e_i^{T_{obs}}; W_d) \tag{8}$$

$$(\Delta x_i^{T_{obs}+1}, \Delta y_i^{T_{obs}+1}) = \delta(d_i^{T_{obs}+1}) \tag{9}$$

where W_d is D-LSTM weight, $\delta(\cdot)$ is a linear layer, $e_i^{T_{obs}}$ is from Eq. 1. The predicted relative coordinate from Eq. 9 at each time-step will be calculated by Eq. 1 to served as input to D-LSTM to infer the relative coordinate of next time-step.

The predicted relative positions are converted to absolute positions for calculating losses. As most previous works [2,6,9], we use a variety loss function that encourages the network to produce diverse samples. For each pedestrian, we generate k possible output predictions by randomly sampling z from $\mathcal{N}(0,1)$ and choosing the "best" prediction in L2 sense as our prediction to compute the loss:

$$L_{variety} = \min_{k} \parallel Y_i - \widehat{Y}_i^k \parallel_2 \tag{10}$$

where Y_i is the ground-truth of future trajectory, $\widehat{Y}_i^k$ is the future trajectory generated by SRGAT, and k is a hyperparameter. To train the network better, only the best trajectory is used to compute the loss to encourage the network to hedge its bets and generate multiple possible future trajectories which are consistent with past trajectory.

3.5 Implementation Details

The dimensions of the hidden state for encoder is 32 and decoder is 80. The input coordinates are embeded as 16 dimensional vectors. The dimension of noise z is set to 16. We use two graph attention layers in RGAT model and the dimensions of intermediate representations is set to 16 and 32 respectively. Adam optimizer with a learning rate of 0.001 is applied to train the model and the batch size is set to 64 in train and test stage.

4 Experiments

We evaluate our method on two public available human walking video datasets: ETH and UCY. These two datasets contain 5 crowd scenes, including ETH, HOTEL, ZARA1, ZARA2, and UNIV. All the trajectories are converted to the world coordinate system and then interpolated to obtain values at every 0.4 s.

Evaluation Metrics. Similar to prior works [6,18], the proposed method is evaluated with two types of metrics as follow:

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.

Baseline. As traditional approaches based on hand-crafted features perform not as well as social LSTM model [1], the traditional models are not listed as baseline. And we only compare SRGAT with the following deep learning based works:

1. *S-LSTM* [1]: An trajectory prediction method that combines LSTM with a social pooling layer, which can aggregate hidden states of the neighbor pedestrians.
2. *SGAN* [6]: An improved version of S-LSTM that the social pooling is displaced with a new pooling mechanism which can learn a "global" pooling vector. A variety loss function is proposed to encourage the GAN to spread its distribution and generate multiple socially acceptable trajectories.
3. *SR-LSTM* [18]: An improved version of S-LSTM by proposing a data-driven state refinement module. The refinement module can jointly and iteratively refines the current states of all participants in the crowd on the basis of their neighbors' intentions through a message passing mechanism.
4. *IA-GAN* [10]: A novel approach to pedestrian prediction that combines generative adversarial networks with a probabilistic model of intent.
5. *TAGCN* [3]: A three stream topology-aware graph convolutional network for interaction message passing between the agents. Temporal encoding of local- and global-level topological features are fused to better characterize dynamic interactions between participants over time.
6. *RSBG* [14]: A novel structure called Recursive Social Behavior Graph, which is supervised by group-based annotations. The social interactions are modeled by GCNs that adequately integrate information from nodes and edges in RSBG.
7. *SRA-LSTM* [13]: A novel social relationship encoder is utilized to learn the social relationships among pedestrians. The social relationship features are added to help modeling social interactions among pedestrians.

Evaluation Methodology. We use the leave-one-out approach similar to that from S-LSTM [1,6]. We train and validate our model on 4 sets and test on the remaining set. We take the coordinates of 8 key frames (3.2 s) of the pedestrian as the observed trajectory, and predict the trajectory of the next 12 key frames (4.8 s).

4.1 Quantitative Evaluations

Table 1 demonstrates the quantitative results between the proposed method and the above mentioned methods across five datasets. The SR-LSTM [18], TAGCN [3], IA-GAN [10], RSBG [14], and our proposed model adopted GNN models (like GCN and GAT) to model social interactions among pedestrians. To our knowledge, only the RSBG [14], SRA-LSTM [13] and our proposed model take social relations into account to model social interactions. On each column, the top three performing methods are highlighted in red, green, and blue. The last column of the table shows the average performance over the five crowd scenes. The SRA-LSTM achieves the minimum ADE and FDE on five dataset. Besides, the SR-LSTM and our proposed model achieve the minimum ADE and FDE respectively.

Due to the difference in motion patterns of pedestrians on each dataset, the performance of a model on the 5 datasets is also different. Thus, we evaluate these

models with a point system on 5 datasets. For each dataset, the top-1, top-2, and top-3 models on ADE or FDE can score 3, 2, and 1 points, respectively. Based on this point system, our proposed model can win 17 points. The SRA-LSTM, RSBG, IA-GAN, TAGCN, SR-LSTM, SGAN, and S-LSTM models can score 14, 9, 8, 2, 12, 6, and 0 points, respectively. According to the scores, our model achieve the best performance on the five datasets.

Table 1. Quantitative results of all the baselines and the proposed method on ETH/UCY datasets. Top-1, top-2, top-3 results are shown in red, green, and blue. (GNN: Graph Neural Networks, SRS: Social Relations)

Method	Notes		Performance (ADE/FDE)					
	GNN	SRS	ETH	HOTEL	ZARA1	ZARA2	UNIV	AVG
S-LSTM	✗	✗	1.09/2.35	0.79/1.73	0.47/1.00	0.56/1.17	0.67/1.40	0.72/1.54
SGAN	✗	✗	0.87/1.62	0.67/1.37	0.35/0.68	0.42/0.78	0.76/1.52	0.61/1.21
SR-LSTM	✓	✗	0.63/1.25	0.37/0.74	0.41/0.90	0.32/0.70	0.51/1.10	0.45/0.94
TAGCN	✓	✗	0.86/1.50	0.59/1.15	0.42/0.90	0.32/0.71	0.54/1.25	0.55/1.10
IA-GAN	✓	✗	0.69/1.42	0.39/0.79	0.35/0.74	0.31/0.66	0.56/1.17	0.46/0.96
RSBG	✓	✓	0.80/1.53	0.33/0.64	0.40/0.86	0.30/0.65	0.59/1.25	0.48/0.99
SRA-LSTM	✗	✓	0.59/1.16	0.29/0.56	0.37/0.82	0.43/0.93	0.55/1.19	0.45/0.93
Ours	✓	✓	0.78/1.59	0.30/0.53	0.35/0.73	0.32/0.66	0.53/1.13	0.46/0.93

Table 2. Parameters size and inference time of different models compared to ours. The lower the better. Models were bench-marked using Nvidia GTX2080Ti GPU. The inference time is the average of several single inference steps. We notice that SRGAT has the least inference time compared to others. The text in blue show how many times our model is faster than others.

	Parameters count	Inference time
SGAN [6]	**46.4K** (0.83x)	0.0057 (1.84x)
SR-LSTM [18]	64.9K (1.17x)	0.0049 (1.58x)
SRA-LSTM [13]	67.1K (1.21x)	0.0045 (1.45x)
SRGAT	55.6K	**0.0031**

Table 2 lists out the speed comparisons between our model and publicly available models which we could bench-mark against. The size of SRGAT is 55.6K parameters. SGAN has the smallest model size with 46.4k parameters, which is about eight tenth of the number of parameters in SRGAT. The sizes of SR-LSTM and SRA-LSTM are 64.9K and 67.1K parameters respectively, which are very close. In terms of inference speed, SRGAT was previously the fastest method with an inference time of 0.0045 s per inference step. However, the inference time of our model is 0.0031 s which is about 1.45x faster than SRA-LSTM.

(a) Parallel Walking (b) People Merging (c) People Meeting

Fig. 3. Comparisons of our method with SRA-LSTM and SR-LSTM in 3 common social scenarios, which containing parallel walking, people merging, and people meeting. The blue solid line represents the observed trajectory, the dashed line represent the future trajectory (blue: ground truth, yellow: SR-LSTM, green: SRA-LSTM, red: our model). (Color figure online)

4.2 Qualitative Evaluations

Benefiting from the proposed social relational graph, SRGAT can learn the latent social relations and better model the social interactions among pedestrians. Thus, SRGAT can perform accurate trajectory prediction. Figure 3 illustrates the trajectory prediction results by using SR-LSTM, SRA-LSTM, and SRGAT in 3 common social scenarios. For the parallel walking cases, the trajectories predicted by SRA-LSTM and SRGAT model are more similar to ground truth. That benefits from the consideration of social relations in models. Furthermore, for more complex social scenarios such as people merging and people meeting, the SRGAT model can still predict the future trajectories which are more similar to the ground truth.

To verify the effect of the proposed model in multimodal trajectory prediction, we compare with the multimodal model SGAN. The multimodal trajectory predictions are shown in Fig. 4. In people meeting scenario, the multimodal trajectories generated by SRGAT are more agminated and tend to avoid the motions of each other. On the contrary, the trajectories predicted by the SGAN model are more dispersed. Similarly, the trajectories predicted by the SRGAT model in people merging scenario are also more agminated. The 3rd column shows the failure case, where neither SGAN nor SRGAT successfully predicted the future trajectories. Since the final destination is unknown, it is difficult to successfully predict the future trajectory of the pedestrian in this case only relying on the 8 time-steps' observed trajectory. That will be the focus of our future work.

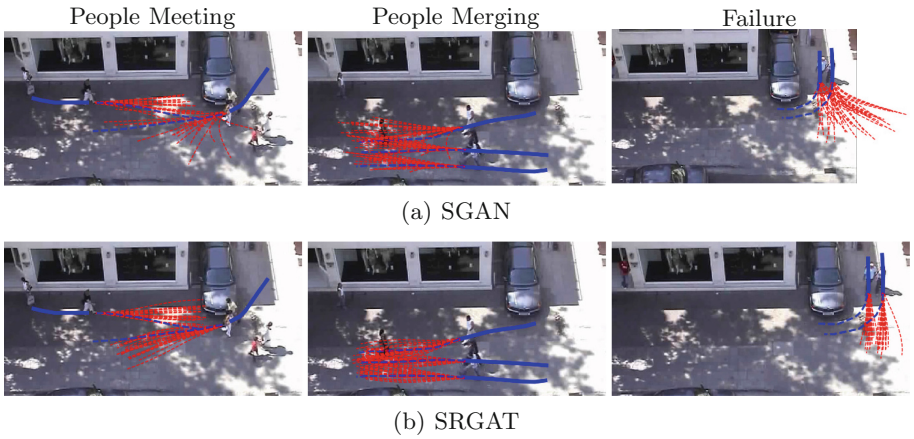

Fig. 4. Comparisons of our method with SGAN in mutimodal trajectory predictions. The first two columns show the results of people merging and people meeting scenarios, and the last column shows a failure case.

5 Conclusions and Discussion

In the work, we designed a social relational graph and modeled the social interactions among pedestrians by relational graph attention network. Two LSTMs were employed to encode the movement of pedestrians and the social interactions among pedestrians to capture the latent motion features and the spatiotemporal interaction features among pedestrians. The encoding features and the random Gaussian noise are fused and then feed to the decoder to generate the multimodal future trajectories. Evaluations are performed in two commonly used metrics, namely, ADE and FDE, across five benchmarking datasets. Comparisons with baseline methods and state-of-the-art approaches indicate the effectiveness of the proposed SRGAT model. The qualitative results of some common social scenarios indicate the success of the use of social relation modeling in trajectory prediction research. To solve the failure cases, we will study the goal guidance of pedestrians and scene information guidance in the future work. In addition, we will study the SRGAT on dynamic interpersonal distance.

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. Alahi, A., Goel, K., Ramanathan, V., et al.: Social LSTM: human trajectory prediction in crowded spaces. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 961–971. IEEE, Las Vegas (2016)
2. Amirian, J., Hayet, J.B., Pettre, J.: Social Ways: learning multi-modal distributions of pedestrian trajectories with GANs. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2964–2972. IEEE, Long Beach (2019)
3. Biswas, A., Morris, B.T.: TAGCN: topology-aware graph convolutional network for trajectory prediction. In: Bebis, G., et al. (eds.) ISVC 2020. LNCS, vol. 12509, pp. 542–553. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-64556-4_42
4. Dan, B., Dane, S., Pietro, C., et al.: Relational graph attention networks. In: International Conference on Learning Representations (ICLR) (2019)
5. Fernando, T., Denman, S., Sridharan, S., et al.: Soft + Hardwired Attention: an LSTM framework for human trajectory prediction and abnormal event detection. Neural Netw. **108**, 466–478 (2018)
6. Gupta, A., Johnson, J., Fei-Fei, L., et al.: Social GAN: socially acceptable trajectories with generative adversarial networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2255–2264. IEEE, Salt Lake City (2018)
7. Hall, E.T.: The Hidden Dimension. Doubleday, Garden City (1966)
8. Helbing, D., Molnar, P.: Social force model for pedestrian dynamics. Phys. Rev. E **51**, 4282 (1998)
9. Huang, Y., Bi, H., Li, Z., et al.: STGAT: modeling spatial-temporal interactions for human trajectory prediction. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6272–6281. IEEE, Seoul (2019)
10. Katyal, K.D., Hager, G.D., Huang, C.M.: Intent-aware pedestrian prediction for adaptive crowd navigation. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 3277–3283. IEEE, Paris (2020)
11. Mohamed, A., Qian, K., Elhoseiny, M., et al.: 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), pp. 14412–14420. IEEE, Seattle (2020)
12. Pellegrini, S., Ess, A., Schindler, K., et al.: You'll never walk alone: modeling social behavior for multi-target tracking. In: 2009 IEEE International Conference on Computer Vision (ICCV), pp. 261–268. IEEE, Miami (2009)
13. Peng, Y., Zhang, G., Shi, J., et al.: SRA-LSTM: social relationship attention LSTM for human trajectory. arXiv preprint arXiv:2103.17045 (2021)
14. Sun, J., Jiang, Q., Lu, C.: Recursive social behavior graph for trajectory prediction. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 657–666. IEEE, Seattle (2020)
15. Vemula, A., Muelling, K., Oh, J.: Social Attention: modeling attention in human crowds. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 4601–4607. IEEE, Brisbane (2018)
16. Xu, Y., Piao, Z., Gao, S.: Encoding crowd interaction with deep neural network for pedestrian trajectory prediction. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5275–5284. IEEE, Salt Lake City (2018)

17. Zhang, M., Liu, X., Liu, W., et al.: Multi-granularity reasoning for social relation recognition from images. In: 2019 IEEE International Conference on Multimedia and Expo (ICME), pp. 1618–1623. IEEE, Shanghai (2019)
18. Zhang, P., Ouyang, W., Zhang, P., et al.: SR-LSTM: state refinement for LSTM towards pedestrian trajectory prediction. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12077–12086. IEEE, Long Beach (2019)
19. Zhang, S., Tong, H., Xu, J., Maciejewski, R.: Graph convolutional networks: a comprehensive review. Comput. Soc. Netw. **6**(1), 1–23 (2019). https://doi.org/10.1186/s40649-019-0069-y