ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
https://doi.org/10.17559/TV-20230727000834
Original scientific paper

Modeling the Uncertainty in Pedestrians Trajectory Prediction

Xiuhong MA, Haitao WANG*, Qiulin MA

Abstract: Pedestrian trajectory prediction attracts a great deal of attention as a fundamental research in the automatic drive system, human-robot interaction, and intelligent surveillance. This paper proposes a spatiotemporal framework based on Transformer and conditional variational autoencoder (CVAE) models to predict pedestrian trajectories while accounting for uncertainty. This paper proposes a spatiotemporal framework based on Transformer and conditional variational autoencoder (CVAE) models to predict pedestrian trajectories while accounting for uncertainty. The Transformer encoder-decoder modules capture spatial and temporal features. The CVAE compares predictions to ground truth across frames to learn motion uncertainty distributions. Experiments on two benchmark datasets demonstrate the approach reduces average displacement error by 0.04 - 0.1 m and final displacement error by 0.05 - 0.15 m compared to prior methods. The results highlight the importance of modeling uncertainty for accurate pedestrian trajectory forecasting.

Keywords: CVAE; pedestrian trajectory prediction; transformer

1 INTRODUCTION

Trajectory prediction tasks are widely used in intelligent surveillance [1]. Robots in a social environment can obtain and analyze motion scenes by monitoring cameras, and make track planning. In an automatic driving system, the motion state of the ego vehicle and others in its driving environment is predicted to be feedback to the system for vehicle driving control. Intelligent monitoring algorithms can assist security personnel in analyzing target movements and detecting abnormal events in a timely manner. Pedestrian trajectory prediction, as one of the techniques for analyzing target movement in video sequences, is regarded as an important foundation for intelligent surveillance applications. Therefore, in order to promote the intelligent development of surveillance video, we improve the pedestrian trajectory prediction task and propose an efficient non-specific scene model. Generally, the dataset of the task is the surveillance video in crowd scenes. The number of pedestrians in the scene is variable, and the pedestrian motion pattern is different. The speed and direction of the individual will change with the neighbors or other factors in the scene. In Figure 1, there are two crowd scenarios being shown, the upper one is a dense crowd in the Zara dataset, and the lower is a sparse crowd in the Hotel dataset. Trajectory prediction aims to predict the future trajectory based on historical motion. Based on the historical trajectories of different pedestrians in different scenes, generalized motion patterns can be learned. At the same time, the prediction model analyses the possible motion interactions between pedestrians and their neighbors and learns the interaction to adjust the pedestrian motion mode. In previous studies [6], these researchers only predict pedestrian trajectories by learning pedestrian motion patterns and interaction relationships, without considering the uncertainty generated by the movement of pedestrians. These models have limitations in uncertainty modeling. For example, the prediction of Social LSTM is a deterministic Gaussian parameter for the next displacement, not the uncertainty of future trajectories. Another method assumes that the future movement distribution obeys the Gaussian distribution when modeling multimodal probability distribution. However, due to the constantly changing motion environment, the movement of the target individual is not only influenced by its own historical trajectory but also constrained by the surrounding environment and neighbor movements, which results in uncertainty in the future direction. Different from this, CVAE mines the future displacement distribution probability by learning historical trajectory features to model multiple possibilities of future trajectories. Sowe analyze the uncertainty generated by pedestrian movement and use the attention-based network to learn the uncertainty feature in movement.

Figure 1 Two pictures of the crowd scenes in Zara and Hotel datasets (Best viewed in color)

We use the attention mechanism to learn the temporal and spatial relationships of the sequences of pedestrian trajectory because the attention mechanism takes into account the different influences between two datasets at different locations when mining the relationships between pairs. Just like that at a certain moment, the interaction between pedestrians in a crowd scene is different. If

influenced by different historical trajectories and path dependency, the same pedestrian's state in the next moment is also different. Therefore, we propose a Transformer [8] based spatiotemporal feature learning framework to model pedestrian motion patterns and interaction relationships. As shown in Figure 2, the pipeline shows the overall process of trajectory prediction of our prediction model. Firstly, learning pedestrian motion patterns by inputting position at $t-1$ and previous historical trajectories. Then, the Conditional Variational AutoEncoder (CVAE) is used to compare the difference between prediction tracks of adjacent frames and ground truth to learn the uncertainty in motion and fine-tune the obtained prediction model parameters. Finally, output the deterministic results and probability prediction results of our model.

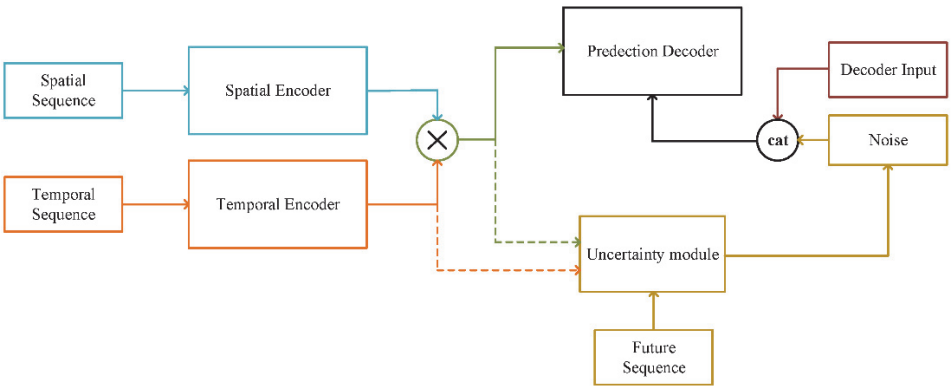

Figure 2 The pipeline of our framework (Best viewed in color)

2 RELATED WORKS

The development of pedestrian trajectory prediction tasks is more and more interesting. Since the team Li proposed Social LSTM [6] in 2016, trajectory prediction models that rely on deep learning have emerged one after another. During this period, researchers used different deep-learning techniques to design various effective prediction models. For example, Social GAN [7], and IDL [9] based on recurrent neural networks use LSTM [10] or GRU [11] networks to learn the temporal characteristics of trajectory. The social-STGCNN [12], SGSG [13], and Graph-TCN [14] methods based on graph neural networks are used to model the spatial relationship of pedestrians in crowd scenes. So Phie [15], and SGSG [13] method based on convolutional neural networks, utilize scene context information to promote the target's understanding of the scene. Researchers propose different prediction modules for different types of information in the learning crowd. To analyze the studies of previous methods in trajectory prediction tasks, we introduce prediction methods into the following two categories from the perspective of feature description: The prediction model of spatiotemporal is based on interaction modeling and motion patterns. Social-LSTM presents a Social pooling layer with a grid map, which also considers the neighborhood around the target. The difference is that they use one LSTM for each trajectory and share the information between the LSTM through the Social pooling layer. Similar to the algorithm used for Social LSTM, researchers usually used LSTM to learn the temporal characteristics of trajectories and modeling spatial coordinate relationships to learn spatial characteristics. Early models such as Social GAN, SR-LSTM [16], and Scene-LSTM [17] all adopt this modeling paradigm. Among them, Social GAN and SR-LSTM use relative coordinates as the basis for global spatial relationship modeling, while Social LSTM and Scene-LSTM divide adjacent regions by occupying the grid. These methods achieve state-of-the-art performance in the early stages of trajectory prediction tasks. For example, the Social-BiGAT [18] method proposes Physical Soft Attention and Social GAT [19] to learn scene features and pedestrian interaction. The STAR [[20] uses a self-attention mechanism to simultaneously learn the temporal and spatial features of pedestrians, and connect spatial transformer and temporal transformer modules by parallel and series for spatiotemporal feature learning. With the deepening of research, researchers find that other information can promote the improvement of prediction model performance to some extent. The motion environment of the target provided, in the scene context, head posture of pedestrian movement, and even road map information. These methods have been used in previous work. In SoPhie, attention modules are established to learn physical attention and social attention for interaction between pedestrians. MX-LSTM [21] uses the head posture to estimate the direction of pedestrian movement, to take into account the relationship between the head deviation angle during pedestrian movement and the current position direction of motion. In summary, the important foundation of trajectory prediction tasks is to model temporal and spatial features. In order to model the spatiotemporal relationship of trajectories, we adopt an attention mechanism-based approach and learn excellent prediction models by analyzing the future uncertainty of different trajectories in the crowd.

3 ATTENTION-BASED SPATIOTEMPORAL MODEL

As shown in Figure 3, we propose a spatiotemporal model based on attention mechanism. Different impacts between individuals are learned to predict trajectories, by modeling trajectories in time and space. Firstly, as a time series, pedestrian trajectories features are learned using an encoder-decoder strategy. Considering different numbers of pedestrians in the scene at the same time, the temporal and spatial input strategies of crowd trajectories is proposed. Then, in the feature decoding stage, the

uncertainty of pedestrian movement is considered, which can result in multiple possibilities for the predicted trajectory in the future. A probability generator CVAE is proposed to learn the uncertainty of spatial and temporal features. Finally, an optimal prediction trajectory is obtained through prediction module.

3.1 Problem Statement

Assuming the current time stamp is t, the historical trajectory of individual i is represented as $X_i^{(1:t-1)}$, where X represents the pedestrian coordinates (x, y), and the future trajectory is represented by $X_i^{(t+1:T)}$, where $T-(t+1)$ is the time length of the future trajectory that needs to be predicted. The index i is the target individual label in the crowd scene, and the number of individuals in each time stamp is defined as N_t.

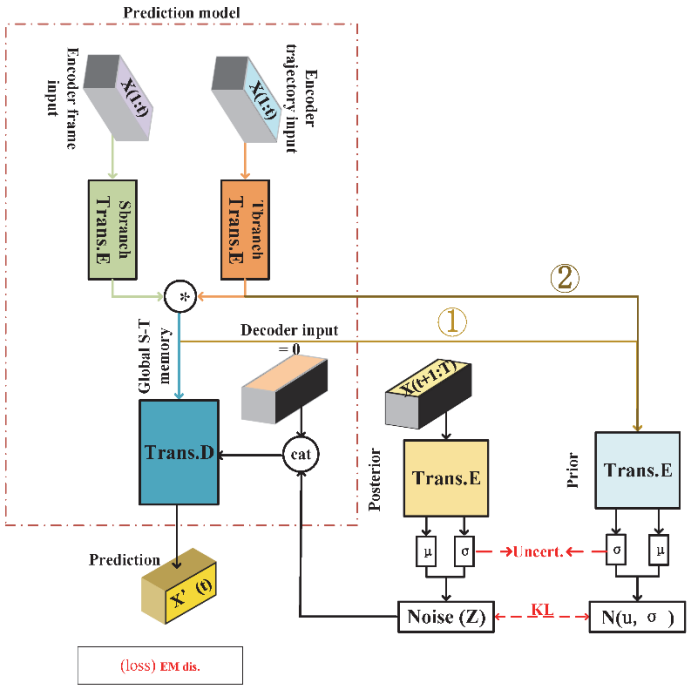

Figure 3 The structure of spatiotemporal branching model. Among them, T branch and S branch represent the temporal and spatial feature encoders of the model, respectively. 1 and 2 represent two different input strategies for the uncertainty learning module. Trans. E and Trans. D in the model is the encoder and decoder modules of theTransformer, and they do not share parameters.(Best viewed in color)

3.2 Spatiotemporal Framework

As mentioned above, our model consists of a spatiotemporal feature encoder-decoder module and an uncertainty learning module. The first part of the model is used to encode and learn the spatial and temporal features of trajectories.

3.2.1 Spatiotemporal Encoder-decoder

Due to the fact that each individual's trajectory is a temporal sequence, we input the trajectory of all pedestrians $X_{i,p}^{(1:t-1)}$ as an independent time series into the temporal feature encoder. Where p represents that the input at this time stamp is a time series. The temporal feature learning formula is as follows:

$$F_t = f_{te}\left(X_{i,p}^{(1:t-1)}\right) \tag{1}$$

where $f_{te}(\cdot)$ is the Transformer encoder module used for learning temporal feature F_t.

At the same time, in order to learn the interaction between different pedestrians, we treat the coordinates of each frame as a spatial sequence $X_{i,s}^{(1:t-1)}$, and input it into the spatial feature encoder to obtain spatial feature where s represents that the input at this time stamp is a spatial sequence. The spatial feature learning formula is as follows:

$$F_s = f_{se}\left(X_{i,s}^{(1:t-1)}\right) \tag{2}$$

where $f_{se}(\cdot)$ is the Transformer encoder module used for learning temporal feature F_s.

As shown in Figure 3, the spatial and temporal encoders are based on the Transformer encoder. By learning sequences of different input formats, the temporal and spatial features of the trajectory are learned separately. Then, the learned spatiotemporal features are fed into the prediction decoder. At the same time, in order to learn the uncertainty of motion, the features are also fed into the uncertainty learning module. The decoder learns that the coordinate formula of the individual in the next time stamp is represented as:

$$\hat{x}_i^{t+1} = f_d\left(\left(F_t \cdot F_s\right), \left[X_i^{(t+1:T)}, Z\right]\right) \tag{3}$$

where the first item ($F_t \cdot F_s$) represents the elements-wise multiplication of temporal and spatial features, and the second item $\left[X_i^{(t+1:T)}, Z\right]$ represents the concatenate between future trajectory and the uncertainty noise Z. This formula is applied to the network training stage.

The predicted coordinate formula during the testing phase is:

$$\hat{x}_i^{t+1} = f_d\left(F_t \cdot F_s\right) \tag{4}$$

3.2.2 Uncertainty Learning Module

The uncertainty learning module analyzes the input temporal or spatial features, which serve as a prior distribution of uncertainty and are encoded through an attention-based encoder. In practice, the uncertainty module uses the encoder layer of the Transformer to learn the parameters of the prior and posterior distributions of the trajectory data. As shown in Figure 3, two different feature input strategies are used as prior, and these two input strategies consider temporal and spatiotemporal features respectively. We analyze the advantages of different input strategies through comparative experiments, demonstrating the importance of spatiotemporal features in uncertainty modeling. As shown in Figure 3, the uncertainty learning model uses future trajectories to learn posterior distribution parameters and calculates the difference between the prior distribution and the posterior distribution variance as the output of the current uncertainty. Moreover, the noise learned from the posterior distribution is used as the distribution of the future trajectories for the prediction model.

3.3 Loss

The loss function is designed according to our model. First, an average displacement loss L_d is designed to calculate the displacement error between the predicted trajectory and the ground truth, which is the prediction error of the model. The CVAE uses KL divergence to approximate the error between the posterior and prior distributions when learning uncertainty in motion and uses it as the training loss L_c of the uncertainty learning module to solve CVAE. The loss function of our model is as follows:

$$L_p = \alpha \cdot L_d + \beta \cdot L_c \tag{5}$$

$$L_d = \left\| \hat{x}_i^{t+1} - x_i^{t+1} \right\|_2 \tag{6}$$

$$\begin{aligned} L_c = & -E_{q_\varphi(Z|X_{t+1:T})}\left[log\ p_\Theta\left(X_{t+1:T}\left|X_{1:t}, Z\right.\right)\right] + \\ & + KL\left(q_\varphi\left(Z\left|X_{t+1:T}\right.\right)\middle\| p_\Theta\left(Z\left|X_{1:t}\right.\right)\right) \end{aligned} \tag{7}$$

where here α and β are the weight parameters of the loss function, $p_\Theta(\cdot)$ is the prior distribution of CAVE and $q_\varphi(\cdot)$ is the posterior distribution.

4 EXPERIMENTS

We use the BIWI [22] and UCY [23] datasets which are the most popular datasets for pedestrian trajectory forecasting tasks. Both BIWI and UCY datasets are derived from high-angle cameras in natural scenes and contain a variety of pedestrian motion patterns, such as going straight, turning toward a destination, or avoiding pedestrians in the middle of the road.

4.1 Experiment Settings

4.1.1 Datasets

The BIWI dataset [22] contains two outdoor scenarios of a long time period. Among these, the eth-university shows an entrance to the school building, and the pedestrians in this scene all have clear destinations. The Hotel shows a street view in front of a Hotel building and the motion of pedestrians is complex in this scene. The UCY [23] is a crowd dataset by UCY Computer Graphics Lab. It includes three different scenarios: Zara Datasets, Arxiepiskopi, and University Students, respectively. The data formats provided by the two above datasets are not uniform. In order to compare them with other benchmark methods, we chose to keep the same with Social LSTM [6] and Social GAN [7], and transformed all data into meters. Furthermore, it is worth noting that, unlike Social GAN, our model predicts deterministic outcomes. Similar to the prior work, we report the prediction error with two metrics:

- Average displacement error (ADE): The mean Euclidean distance error over all estimated points of a trajectory and the true points.

$$E_{ADE} = \frac{1}{N}\frac{1}{T-t}\sum_{i=1}^{N}\sum_{k=t+1}^{T}\sqrt{\sum_{n=1}^{2}\left(x_k^{i,n} - \hat{x}_k^{i,n}\right)^2} \tag{8}$$

- Final Displacement error (FDE): The distance between the predicted final destination and the true final destination at the end of the prediction period T.

$$E_{FDE} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\sum_{n=1}^{2}\left(x_T^{i,n} - \hat{x}_T^{i,n}\right)^2} \tag{9}$$

Following [6], we use the leave-one-out approach, where four of the five scenes are used for training and validating, and the remaining one is used for testing. In these experiments, we set the sampling interval of videos in BIWI and UCY to 10 frames. In other words, the interval is 0.4 secs at every time step. To be specific, the observed length of the trajectory is 8 time steps (3.2 secs) and shows prediction results for 12 time steps (4.8 secs).

4.1.2 Implementation Details

Two-dimensional coordinates as the input are embedded into a vector in size of 64. All of the transformer layers accept input with feature size 64. The dropout ratio of 0.1 is applied when processing the encoded data. Both

spatial and temporal branch transformer consists of 6 encoding layers with 8 heads for each. In addition, we train the network using Adam optimizer with a learning rate of 0.1 and batch size 18 for 120 epochs. Each batch contains the 12 time steps with different numbers of pedestrians indicated by an attention mask to accelerate the training and inference process.

4.2 Ablation Experiment

To verify the performance of the uncertainty learning module proposed in Figure 2, different input strategies are set for the module, and ablation experiments are conducted to analyze the impact of different strategies on model performance. In Table 1, the Initial noise-based method represents the initialized noise as the uncertainty module input, which does not encode future trajectories with CVAE. This strategy assumes that the probability of the future trajectory distribution of the pedestrian follows a Gaussian distribution. The Uncertainty-based 1 method uses spatiotemporal features learned from the spatiotemporal model as input to the prior distribution of the uncertainty module, and inputs future trajectories as learning information to the CVAE posterior distribution encoder. The difference between the Uncertainty-based 2 and 1 method is that it uses the motion features learned from the temporal encoder as input to the prior distribution decoder, while other conditions remain unchanged.

Table 1 The ablation experimental results of the uncertainty learning module

Methods	Metrics	Eth	Hotel	Zara1	Zara2	Univ
Initial noise-model	ADE	0.653	0.382	0.595	0.581	0.611
	FDE	1.123	0.693	1.095	1.069	1.113
Uncertainty-based 1	ADE	0.646	0.410	0.579	0.595	0.624
	FDE	1.110	0.747	1.069	1.095	1.136
Uncertainty-based 2	ADE	0.647	0.413	0.580	0.603	0.620
	FDE	1.111	0.753	1.069	1.109	1.129

From the experimental results in Table 1, it can be seen that although the prediction results of the three models mentioned above achieved the best in different real crowd scene, respectively. However, the Uncertainty-based 1 method outperforms the Uncertainty-based 2 method in most scenes. There are slight differences in the Univ scene, which may be due to the fact that the population density in this scene is much higher than in other scenes, making the model unable to predict pedestrian interaction relationships well. The prediction performance of the Initial noise-based method is significantly better than the other two methods on Univ sequences, indicating that the uncertainty model has a disadvantage in learning the interaction relationships between dense populations. The performance of the Uncertainty-based 2 method is better than that of the Initial noise-based in Zara1 sequences with clearer group movements, while the performance is opposite in Zara2 sequences with more complex group movements. This result further confirms the above conclusion. The results of ablation experiments indicate that our model has shortcomings in modeling dense populations when learning interaction relationships within the population. In future model improvements, greater emphasis should be placed on the ability to model complex interactive changes in densely populated populations.

4.3 Results Analysis

We use 5 video datasets to conduct experiments, and the experimental results are quite different in different video datasets. As shown in

Table 2, our model obtains the best results on the Eth dataset, and the optimal FDE error on Zara2 and Univ. The ADE error on Hotel and Univ is very close to the optimal result. The results show that our method outperforms the benchmark method Social LSTM and outperforms Social GAN in three videos, and also achieves the same level of performance as Social GAN on the other two datasets. This proved the effectiveness of our proposed model.

Table 2 Trajectory prediction results on 5 datasets

Methods	Metrics	Eth	Hotel	Zara1	Zara2	Univ
Ours Uncertainty- based 1	ADE	0.64	0.41	0.57	0.59	0.62
	FDE	1.11	0.74	1.06	1.09	1.13
Social GAN(1v-1)	ADE	1.13	1.01	0.42	0.52	0.60
	FDE	2.21	2.18	0.91	1.11	1.28
Social LSTM	ADE	1.09	0.79	0.47	0.56	0.67
	FDE	2.35	1.76	1.00	1.17	1.40
LSTM	ADE	1.09	0.86	0.41	0.52	0.61
	FDE	2.41	1.91	0.88	1.11	1.31
Linear	ADE	1.33	0.39	0.62	0.77	0.82
	FDE	2.94	0.72	1.21	1.48	1.59
Social Attention	ADE	1.39	2.51	1.25	1.01	0.88
	FDE	2.39	2.91	2.54	2.17	1.75

Tab. 2 reflects that our method achieved optimal results in some scenes. In addition, the reasons for the decline in model prediction performance in other scenarios are analyzed as follows. In the scene of Hotel video, the Linear method achieves the best results, while our method achieves suboptimal results, indicating that the pedestrian motion patterns in the video were simple. Using the attention mechanism to model motion features obtained a better model than using LSTM, demonstrating the effectiveness of the attention mechanism in modeling motion features of time sequence. In the same scene of Zara1 and Zara2, the LSTM-based method performs better, and the LSTM model without interaction modeling performs best. This indicates that the moving target in this scene is sensitive to modeling pedestrian interaction relationships, and interaction modeling to some extent affects the predictive performance of the model.

5 CONCLUSION

We propose a trajectory prediction model based on uncertainty modelling, which uses CVAE to determine the uncertainty of the future displacement. Trajectory motion pattern and the interaction characteristics between pedestrians are extracted with Transformer, an attention based model. Our approach has a good effect on some datasets, but there are limitations in other datasets. Among them, the analysis for affecting the predictive performance is as follows: when the crowd density is moderate and pedestrian interaction is simple, the model's predictive performance is better than that of the dense and complex interaction scenarios. This indicates it is necessary to consider more comprehensively the external interaction effects of pedestrians from dense crowd. For example, the impact of small groups or high-density crowd on pedestrian movement. The proposed method can be utilized for intelligent security monitoring, such as ensuring the spatial distance between tourists and exhibits in museums, or for intelligent driving by assisting drivers in assessing the movement trajectory of pedestrians ahead of the vehicle to prevent collisions, among other applications. In future research, we will also learn the impact of different pedestrian densities on predicting pedestrian trajectories in different scenes. In addition to the motion patterns and interactions of pedestrians, additional considerations such as the characteristics of the crowd, the flow of people, and the structure of the scene make the research on pedestrian trajectory prediction tasks more comprehensive.

Acknowledgements

This work is supported by the Fundamental Research Project of Humanities and Social Sciences in Colleges and Universities of Hebei Province (NO. GH171046). We would like to show our deepest gratitude to the colleague, Ming Chen, who has provided us with valuable guidance in every stage of the writing of this paper. His support includes the following aspects: proper English language, grammar, punctuation, and spelling.

6 REFERENCES

[1] Zolghadr, J. & Cai, Y. (2015). Locating a two-wheeled robot using extended Kalman filter. *Tehnički vjesnik, 22*(6), 1481-1488. https://doi.org/10.17559/TV-20140531190647

[2] Qiu, Y. & Liu, C. (2014). Modelling and stimulation of target tracking and localization in wireless sensor network. *Tehnički vjesnik, 21(*2), 233-238.

[3] Chen, Z., Zhang, Y., Wu, C., & Ran, B. (2019). Understanding Individualization Driving States via Latent Dirichlet Allocation Model. *IEEE Intelligent Transportation Systems Magazine, 11*(2), 41-53. https://doi.org/10.1109/MITS.2019.2903525

[4] Qiao, S., Tang, C., Jin and H., Long, T., Dai, S., Ku, Y., & Chau, M. (2010). PutMode: prediction of uncertain trajectories in moving objects databases. *Applied Intelligence, 33*, 370-386. https://doi.org/10.1007/s10489-009-0173-z

[5] Chen, Z., Cai, H., Zhang, Y., Wu, C., Mu, M., Li, Z., & Sotelo, M. A. (2019). A novel sparse representation model for pedestrian abnormal trajectory understanding, *Expert Systems with Applications, 138*. https://doi.org/10.1016/j.eswa.2019.06.041

[6] Alahi, A., Goel, K., Ramannathanand, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). Social LSTM: Human Trajectory Prediction in Crowded Spaces. *IEEE Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR.2016.110

[7] Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., & Alahi, A. (2018). Social GAN: Socially acceptable trajectories with Generative adversarial networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2018.00240

[8] Vaswani, A., Shazeer, N., and Parmar, N., and Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017) Attention is all you need. *International Conference on Neural Information Processing Systems (NIPS'17)*, 6000-6010.

[9] Li, Y. (2019) Which Way Are You Going? Imitative Decision Learning for Path Forecasting in Dynamic Scenes. *IEEE Conference on Computer Vision and Pattern Recognition*, 294-303. https://doi.org/10.1109/CVPR.2019.00038

[10] Hochreiter, S. & Schmidhuber, J. (1997) Long short-term memory. *Neural Computation, 9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[11] Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724-1734. https://doi.org/10.3115/v1/D14-1179

[12] Mohamed A, Qian K, Elhoseiny M., & Claudel, C. (2020) Social-STGCNN: A social spatio-temporal graph convolutional neural network for human trajectory prediction. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14412-14420. https://doi.org/10.1109/CVPR42600.2020.01443

[13] Xue, H., Huynh, D. Q., & Reynolds, M. (2020) Scene Gated Social Graph: Pedestrian Trajectory Prediction Based on Dynamic Social Graphs and Scene Constraints. *Computer Vision and Pattern Recognition*.

[14] Wang, C., Cai, S., & Tan, G. (2021) GraphTCN: Spatio-Temporal Interaction Modeling for Human Trajectory Prediction. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 3449-3458. https://doi.org/10.1109/WACV48630.2021.00349

[15] Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Rezatofighi, H., & Savarese, S. (2019). SoPhie: An Attentive GAN for Predicting Paths Compliant to Social and Physical

Constraints. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2019.00144
[16] Zhang, P., Ouyang, W., Zhang, P., & Xue, J. (2019). SR-LSTM: State Refinement for LSTM towards Pedestrian Trajectory Prediction. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2019.01236
[17] Manh, H. & Alaghband, G. (2018). Scene-LSTM: A Model for Human Trajectory Prediction.
[18] Kosaraju, V., Sadeghian, A., Martin-Martin, R., Reid, I., Rezatofighi, H., & Savarese, S. (2019) Social-BiGAT: Multimodal Trajectory Forecasting using Bicycle-GAN and Graph Attention Networks.
[19] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *The International Conference on Neural Information Processing Systems, 2*, 2673-2680.
[20] Yu, C., Ma, X., Ren, J., Zhao, H., & Yi, S. (2020). Spatio-Temporal Graph Transformer Networks for Pedestrian Trajectory Prediction. https://doi.org/10.1007/978-3-030-58610-2_30
[21] Hasan, I., Setti, F., Tsesmelis, T., Belagiannis, V., Amin, S., Del Blue, A., Cristani, M., & Galasso, F. (2021) Forecasting people trajectories and head poses by jointly reasoning on tracklets and vislets. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 43*(4), 1267-1278. https://doi.org/10.1109/TPAMI.2019.2949414
[22] Pellegrini, S., Ess, A., Schindler, K., & van Gool, L. (2009). You'll never walk alone: Modeling social behavior for multi-targer tracking. *IEEE 12th International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/ICCV.2009.5459260
[23] Lerner, A., Chrysanthou, Y., & Lischinski, D. (2007). Crowds by example. *Computer Graphics forum, 26*(3), 655-664. https://doi.org/10.1111/j.1467-8659.2007.01089.x

Contact information:

Xiuhong MA, PhD
Hebei University of Economics and Business,
School of Management Science and Engineering,
Hebei University of Economics and Business, Shijiazhuang, 050061, China
E-mail: maxiuhong@heuet.edu.cn

Haitao WANG, senior engineer
(Corresponding author)
Hebei University of Economics and Business,
Modern Educational Technology Center,
Hebei University of Economics and Business, Shijiazhuang, 050061, China
E-mail: wanght@heuet.edu.cn

Qiulin MA, PhD student
Beijing Jiaotong University,
Beijing Key Laboratory of Traffic Data Analysis and Mining,
Beijing Jiaotong University, Beijing, 100044, China
E-mail: 17112075@bjtu.edu.cn