REGION: Relevant Entropy Graph spatIO-temporal convolutional Network for Pedestrian Trajectory Prediction

Naiyao Wang[1], Yukun Wang[1], Changdong Zhou[1], Ajith Abraham[2], and Hongbo Liu[1(✉)]

[1] College of Information Science and Technology, Dalian Maritime University, Dalian, Liaoning, China
{wny,lhb}@dlmu.edu.cn
[2] Machine Intelligence Research Labs (MIR Labs), Auburn, WA, USA

Abstract. Modeling pedestrian interaction is an essential building block in pedestrian trajectory prediction, which raises various challenges such as the complexity of social behavior and the randomness of motion. In this paper, a new relevant entropy spatio-temporal graph convolutional network is proposed to model pedestrian interaction for pedestrian trajectory prediction, which contains regional spatiotemporal graph convolutional neural network and gated dilation causal convolutional neural network. The regional spatio-temporal graph convolutional neural network creates a matching graph structure for each time step, and calculates the weighted adjacency matrix of each graph structure through relevant entropy to obtain the sequence embedding representation of the pedestrian interaction relationship. The gated dilation causal convolutional neural network reduces the linear superposition of the hidden layer through the setting of the dilated factor, and uses the gating mechanism to filter the features. Experiments are carried out on the standard data sets ETH and UCY, higher accuracy and efficiency verify that the proposed method is effective in pedestrian interaction modeling.

Keywords: Graph convolution · Relevant entropy · Gating mechanism · Trajectory prediction

1 Introduction

As an important participant in traffic scenes, pedestrians exhibit highly random movements [1,2]. Predicting the trajectory of pedestrians is of great significance in many traffic fields such as automatic driving and monitoring systems [3,4]. Pedestrians are subjective when making route decisions, and common sense of social rules needs to be followed [5,6]. The motion subject needs to analyze other human's actions and social behaviors to adjust their own routes. In addition, there exist interactions in the group environment, and individual behavior patterns will be implicitly affected by the surrounding environment [7–9]. Therefore,

A. Abraham et al. (Eds.): IBICA 2021, LNNS 419, pp. 150–159, 2022.
https://doi.org/10.1007/978-3-030-96299-9_15

building a pedestrian interaction model with high interpretability and generalization is the focus of the trajectory prediction problem [10,11].

At present, data-driven methods regard pedestrian trajectory prediction as a time series data prediction problem. Although certain progress has been made, there are still some difficulties needed to be resolved, which can be summarized as the following aspects: 1) The interpretability of the pedestrian trajectory prediction problem is poor. The high-dimensional information in the scene is difficult to fit, the physical meaning is difficult to interpret, and it is not intuitive to model pedestrian scene information. 2) Pedestrian interaction is difficult to model. Pedestrian interaction is subjective. Because it changes dynamically with the transformation of time and space and is affected by potential social rules, the scope of interaction is difficult to define. 3) Pedestrian trajectory prediction is a long-sequence prediction problem, with a large amount of data calculation and a lot of interference information, which is difficult to fit the nonlinear relationship in the time dimension.

In this work, we propose a new regional relevant entropy spatiotemporal graph scene modeling method, and on this basis, a pedestrian trajectory prediction model based on graph convolutional neural network is presented. The main body model consists of two main parts: the regional relevant entropy spatiotemporal graph convolutional neural network and the gated dilation causal convolutional neural network. In the regional relevant entropy spatio-temporal graph, the relevant entropy is introduced to calculate the weight of each edge and the weighted adjacency matrix indicates the strength of the mutual influence between pedestrians, then the feature of the pedestrian's past trajectory is extracted from the matrix by means of convolution operation. By taking the obtained features as input, the gated dilation causal convolutional neural network operates on the time dimension of the embedding result, which reduces the error of the model through a gating mechanism, and finally predicts the multimodal trajectory of all pedestrians in the future.

The main contributions of this paper are listed below:

- We propose a new regional relevant entropy graph spatio-temporal convolutional network for modeling pedestrian groups in traffic scenes, called the REGION model. The topology of the graph is a natural way to represent the social interaction between the pedestrians in the scene.
- A new gated dilation causal convolutional neural network is proposed for time series prediction. It can effectively filter useless features and prevent the problem of gradient disappearance during training.
- The proposed model is trained on the standard public data set, and the obtained model is compared with other baseline methods on data and visualization.

2 Related Work

Physics-based model predicts movement by simulating a set of well-defined dynamics equations. Zernetsch et al. use a physical model of cyclists containing the driving and resistance forces to predict their future position [12]. Kooij et al. address the problem of predicting the path of objects with multiple dynamic modes by the introducing of the latent variables related to pedestrian awareness [13]. Following the success of Recurrent Neural Network models for sequence prediction tasks, a Long Short-term Memory model which can learn general human movement and predict their future trajectories is proposed in the work of Alahi et al. [14]. Later, the combining of generative adversarial networks provides a new idea to solve the problem of sequence prediction, Gupta et al. propose a recurrent sequence-to-sequence model observes motion histories and predicts future behavior [15]. With the development of graph convolution technology, Mohamed et al. propose spatiotemporal graph convolutional network and temporal extrapolation network to extract features by spatiotemporal convolution on the representation of pedestrian trajectories [16]. The kind of model makes predictions by specifying motion goals and formulating strategies to achieve them. Shen et al. develops a transferable pedestrian motion prediction algorithm based on Inverse Reinforcement Learning (IRL) that infers pedestrian intentions and predicts future trajectories based on observed trajectory [17]. Further research proposes a new general framework for directly extracting a policy from data, as if it is obtained by reinforcement learning following inverse reinforcement learning [18]. By minimizing the symmetrized cross-entropy between the distribution and demonstration data, Rhinehart et al. proposes a method to forecast a vehicle's ego-motion as a distribution over spatiotemporal paths, conditioned on features embedded in an overhead map [19]. Xie et al. proposes a sequential model that combines the CNN with the LSTM to predict surrounding vehicle trajectories [20]. Zhao et al. proposes a CNN-based model for human pedestrian trajectory prediction with the idea of motion patterns [21]. In order to model the interaction between pedestrians in the scene more reasonably while improving the efficiency of the model, this paper adopts a relevant entropy graph spatio-temporal convolutional network.

3 Methodology

The overall framework of the proposed REGION is presented in Fig. 1. First, take the position of each pedestrian in the dataset that has been marked into the model and construct the regional spatial map structure: the position information of the pedestrian is used as the nodes set in the map structure, the distance information between the two is used as the edge attribute between their nodes if the distance between pedestrians is within the threshold set in advance. The edge set in the graph structure indicates the influence between pedestrians; Secondly, for any node in the graph, calculate the influence relationship coefficient between the two directly adjacent nodes through the relevant entropy and the

weighted adjacency matrix of the graph structure can be formed; Then, expand the regional space map into a regional space-time map and perform convolution operations on the weighted adjacency matrix of the regional spatio-temporal graph at each time for getting the embedding representation of the graph at different times; At last, the obtained embedding result is put into the gated dilation causal convolutional neural network, which predicts on the time dimension to obtain the final trajectory prediction result.

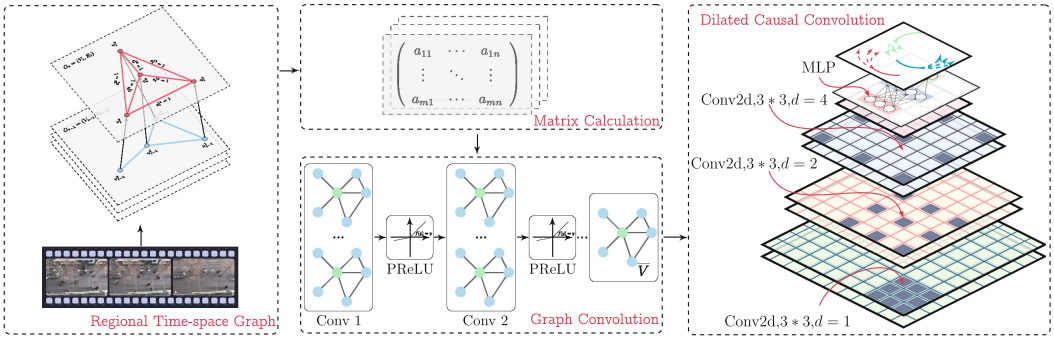

Fig. 1. The overall framework of the proposed REGION.

Since the space-time graph changes gradually with time steps, before constructing the space-time map, this paper first constructs a set of space maps G_t to represent the position of the pedestrian in the scene at any time t in the scene. Given a topological graph $G_t =< V_t, E_t, R_t >$, where $V_t = \{v_t^i | i \in (1, ..., N)\}$ is the vertex set, $E_t = \{e_t^{ij} | i \in (1, ..., N)\}$ is the edge set and $R_t = \{r_t^{ij} | i \in (1, ..., N)\}$ is the nodes relevance set.

Let $S \in \{s_1, s_2, ..., s_n\}$ be the scope of pedestrian interaction at time t, $R_t \in \{r_1^{ij}, r_2^{ij}, ..., r_m^{ij}\}$ be the nodes relevance of vertexes v_t^i and v_t^j at time t, if given S, the relevant entropy of R_t is defined as Eq. (1):

$$a_t^{ij} = H(R_t|S) = -\sum_{i,j=1}^{m} p(r_t^{ij}|S = \varepsilon) \log p(r_t^{ij}|S = \varepsilon). \tag{1}$$

The calculation method of $R_t \in \{r_1^{ij}, r_2^{ij}, ..., r_t^{ij}\}$ needs to consider the distance and direction of pedestrians. We use cosine similarity to calculate it as Eq. (2):

$$r_t^{ij} = \frac{v_t^i \cdot v_t^j}{||v_t^i|| ||v_t^j||}. \tag{2}$$

All the a_t^{ij} constitutes a weighted adjacency matrix A, then we need to normalize the matrix A. Through this operation, the data can be made comparable and the relationship between the data can be relatively maintained. We need to

multiply $A \in \{A_1, A_2, ..., A_t\}$ by the degree matrix D^{-1} and further divide it into two $D^{-\frac{1}{2}}$, to obtain a symmetric and normalized matrix as Eq. (3):

$$A_t = D^{-\frac{1}{2}} * \widehat{A_t} * D^{-\frac{1}{2}}. \tag{3}$$

where $\widehat{A_t} = A + \beta I_N$, taking $\beta = 1$ (makes the characteristics of the node itself as important as its neighbors), we have $\widehat{A_t} = A + I$, and I is the identity matrix.

Then, the aggregate information around the target node is extracted through the convolution operation on the graph structure. The convolution operation definition of the regional spatiotemporal graph convolutional neural network can be obtained, as shown in Eq. (4):

$$v^{i(l+1)} = \sigma(\lambda \sum_{v^j \in \tau} P(\kappa^{(l)}) \cdot W(v^{i(l)}, v^{j(l)})). \tag{4}$$

where σ is the PRelu activation function, λ is a normalization item, and $P(\cdot)$ represents the sampling process, which aggregates the surrounding information centered on κ. (l) represents the l layer, and $W(\cdot)$ represents the weight that needs to be trained in the network, $\tau = \{v^j | d(v^i, v^j) \leq D\}$ is the set of adjacent vertices of the vertex v^i. $P(\kappa^{(l)})$ aggregates the embedded information of neighbor nodes and its own node. We calculate $P(\kappa^{(l)})$ as Eq. (5):

$$\kappa^{i(l)} = \psi_l(v^{i(l-1)}, \mu_l(v^{j(l-1)}, \forall j \in S)). \tag{5}$$

where $\mu_l(\cdot)$ is the aggregate function, and $\psi_l(\cdot)$ is the concat function. In summary, the equation of Regional spatio-temporal graph convolutional neural network is shown in Eq. (6).

$$R(V^{(l)}, A_t) = \sigma(A_t V^{(l)} W^{(l)}). \tag{6}$$

where $W^{(l)}$ represents the matrix of the trainable parameters of the l layer. After applying the graph convolutional neural network, the features of the graph that can be represented compactly are obtained. The embedding result obtained is expressed as $\widetilde{V}$.

Time series prediction requires that the prediction result at a certain time t can only be judged by the input before time t, this paper proposes a gated dilation causal convolutional neural network.

The paper defines filter $f : \{0, 1, ..., k-1\} \to \mathbb{R}$, the input sequence is $\widetilde{V} = (\widetilde{v}_1, \widetilde{v}_2, ..., \widetilde{v}_t)$, it is the embedding result sequence output by the spatiotemporal graph convolutional neural network. The convolution kernel with the dilation factor d at the sequence s is shown in Eq. (7).

$$F(s) = \sum_{i=0}^{k-1} f(i) \cdot \widetilde{v}_{s-di}. \tag{7}$$

where k represents the size of the convolution kernel. In this paper, the dilation factor of the input layer is defined as 1, which is the ordinary convolution.

The dilation factor of the second hidden layer is set to 2. As the network layer increases, the expansion factor is 2 increases in exponential form.

In order to get a better training effect on the nonlinear relationship in the time dimension, this module sets an output gate on the convolutional layer of the dilated causal convolution, and sets independent training parameters for it. The output of each hidden layer is controlled by the output gate. Carry out regulation. The calculation formula for each hidden layer can be expressed as Eq. (8).

$$h^l = \varphi(\gamma_1 X_d + b) \otimes \delta(\gamma_2 X_d + c). \tag{8}$$

where $\delta(\gamma_2 X_d + c)$ is the output gate of the convolutional layer, through which the mapping $\varphi(\gamma_1 X_d + b)$ can be adjusted to make the weight is more suitable for the prediction model. Where φ and δ represent the activation function, respectively, γ_1 and γ_2 are the weights required to train the model, X_d is the feature in the expanded causal convolution, $\otimes$ represents the element product operation between the matrices, b and c are bias term.

4 Experiments

In this section, we will introduce experiments to verify the validity of the model. The training efficiency of this model is compared with other methods, which proves the efficiency of this method. And the prediction accuracy and visualization are compared with this model and multiple baseline methods, such as SocialLSTM [14], SocialGAN [15] and SoPhie [22].

The paper uses the average displacement error and the final displacement error to quantitatively analyze the model. All models take 8 frames as input and make predictions for the next 12 frames. The results of the prediction model are shown in Table 1 and Fig. 2.

Table 1. ADE/FDE results on trajectory prediction of different methods.

Datasets	Linear [14]	Sophie [22]	S-GAN [15]	S-LSTM [14]	REGION
eth	1.35/2.96	0.72/1.45	0.85/1.64	1.10/2.37	**0.66/1.13**
hotel	**0.37/0.69**	0.75/1.65	0.70/1.39	0.78/1.74	0.51/0.88
zara1	0.59/1.19	**0.33**/0.63	0.34/0.67	0.45/0.98	0.36/**0.59**
zara2	0.75/1.49	0.39/0.76	0.42/0.85	0.52/1.15	**0.30/0.61**
univ	0.83/1.60	0.52/1.22	0.75/1.50	0.67/1.42	**0.47/0.83**
avg	0.78/1.59	0.54/1.14	0.61/1.21	0.70/1.53	**0.46/0.80**

It can be seen from the Table 1 and Fig. 2 that the REGION model studied in this paper constructs a regional spatio-temporal graph and calculates the weighted adjacency matrix on the graph through relevant entropy, so that the

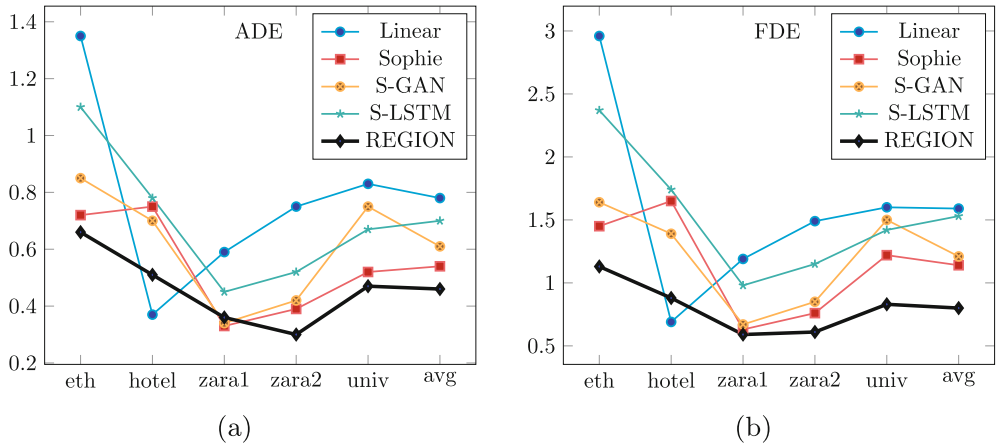

Fig. 2. Trajectory evaluation ADE/FDE results on pedestrian trajectory prediction. (a) ADE value of the evaluation results; (b) FDE value of the evaluation results.

effective social interaction information around the target pedestrian is adopted by the network and the interference information is discarded, which has obtained better prediction accuracy on the standard data set, and can accurately predict the movement trend of pedestrians in the future.

The paper uses a pre-divided test set to test the model, and provides the visualization of the trajectory prediction results of four scenarios: pedestrians from different directions merging together, pedestrians walking in opposite directions, pedestrians walking side by side in the same direction, and a single pedestrian meeting a group of pedestrians. In view of the above four scenarios, the paper compares the visualization results on trajectory prediction of the SoPhie model, the Social-GAN model and the REGION model in Fig. 3.

We also show the visualization results on the multi-modal trajectory predictions of the different model, as shown in Fig. 4, which includes situations where pedestrians catching up in the same direction, individual pedestrians encountering groups in different directions, converge walking, and waiting for turning. After analyzing the evaluation results, it is proved that the REGION model proposed in this paper can predict the multiple walking routes of multiple pedestrians in the scene for a period of time in the future, and the overall prediction trend is consistent with the actual walking path of pedestrians.

It can be seen from Fig. 4. SoPhie and REGION give a collision-free path prediction. However, through the visualization results, it can be intuitively observed that the accuracy of the prediction results of the REGION model proposed in this paper is slightly higher than that of the other models, and the path is more similar to the real trajectory of pedestrians.

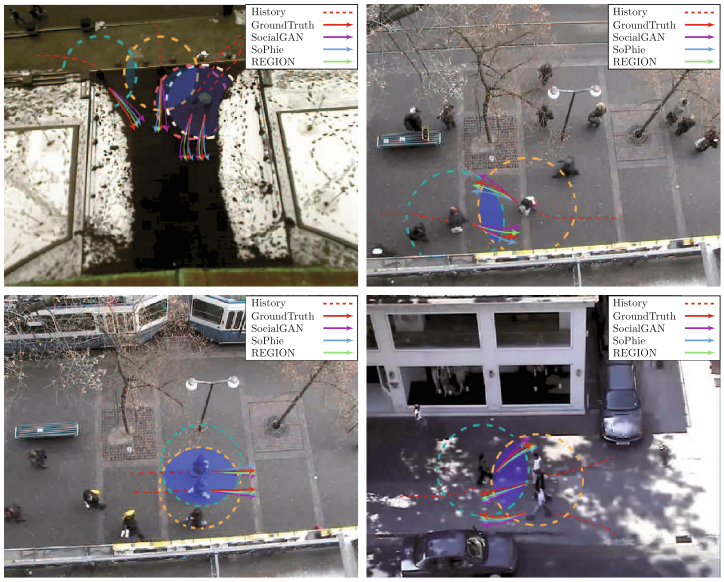

Fig. 3. Trajectory prediction accuracy visualization results of different methods. The dashed circle represents the regional influence range of each pedestrian. The deeper the blue at the intersection of the two circles is, the greater the degree of influence between the pedestrians will be.

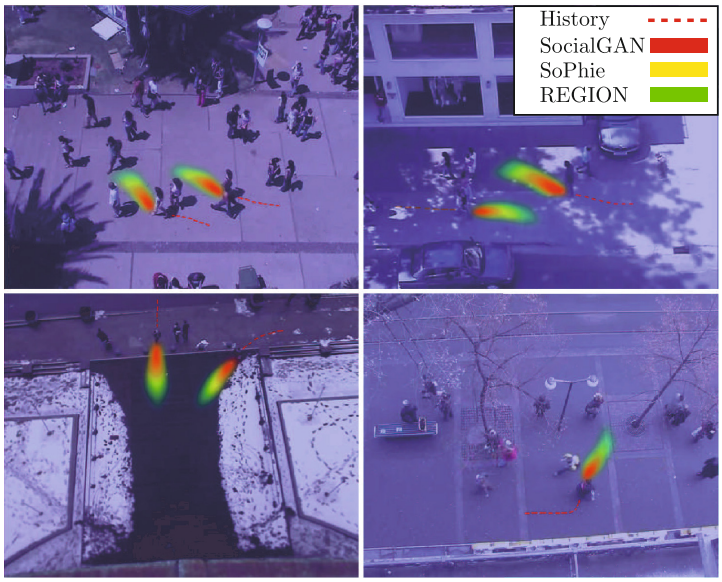

Fig. 4. Visualization of the multi-modal trajectory predictions on different methods.

5 Conclusions

The main research content of this paper is to complete the prediction of the trajectory when a group of pedestrians interact with each other under a fixed monitoring perspective. A regional graph spatio-temporal convolutional neural

network algorithm is proposed. The model defines the conditions for the connections between nodes when constructing the regional spatio-temporal graph, so that the modeling of pedestrian groups is more suitable for the actual situation in society. A gated dilation causal convolutional neural network is also proposed, which can obtain good prediction results while ensuring training efficiency. In addition, the dilation of causal convolution allows the receptive field to increase simultaneously with the increase in network depth, so that the model can make accurate predictions of pedestrians' future walking trajectory trends. The model REGION proposed in this paper generally has a good performance on the data set with moderate crowd density in the scene, but there are errors in the prediction of the data set environment with a single scene and less nonlinear relationship between pedestrians. In the follow-up work, we should adopt different modeling strategies to predict the future trajectory in different complex scenarios, entropy-based energy models will be used to more accurately model the interaction between pedestrians, and transformer-based graph convolutional networks can be introduced to better extract pedestrian features.

Acknowledgements. This work is supported in part by the National Natural Science Foundation of China (Grant Nos. 61772102, 62176036) and the Liaoning Collaborative Fund (Grant No. 2020-HYLH-17).

References

1. Zhao, H., Wildes, R.P.: Where are you heading? Dynamic trajectory prediction with expert goal examples. In: Proceedings of the International Conference on Computer Vision, pp. 7629–7638 (2021)
2. Peng, Y., Zhang, G., Li, X., Zheng, L.: STIRNet: a spatial-temporal interaction-aware recursive network for human trajectory prediction. In: Proceedings of the International Conference on Computer Vision, pp. 2285–2293 (2021)
3. Cai, Y., et al.: Pedestrian motion trajectory prediction in intelligent driving from far shot first-person perspective video. IEEE Trans. Intell. Transp. Syst. (2021, in press). https://ieeexplore.ieee.org/document/9340008
4. Quan, R., Zhu, L., Wu, Y., Yang, Y.: Holistic LSTM for pedestrian trajectory prediction. IEEE Trans. Image Process. **30**, 3229–3239 (2021)
5. Li, Y., Liang, R., Wei, W., Wang, W., Zhou, J., Li, X.: Temporal pyramid network with spatial-temporal attention for pedestrian trajectory prediction. IEEE Trans. Netw. Sci. Eng. (2021, in press). https://ieeexplore.ieee.org/document/9373939
6. Cai, Y., et al.: Environment-attention network for vehicle trajectory prediction. IEEE Trans. Veh. Technol. **70**, 11216–11227 (2021)
7. Pang, B., Zhao, T., Xie, X., Wu, Y.N.: Trajectory prediction with latent belief energy-based model. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition, pp. 11814–11824 (2021)
8. Zhang, B., Yuan, C., Wang, T., Liu, H.: STENet: a hybrid spatio-temporal embedding network for human trajectory forecasting. Eng. Appl. Artif. Intell. **106**, 104487 (2021)
9. Chen, G., Li, J., Lu, J., Zhou J.: Human trajectory prediction via counterfactual analysis. In Proceedings of the International Conference on Computer Vision, pp. 9824–9833 (2021)

10. Zhang, B., Zhang, R., Bisagno, N., Conci, N., De Natale, F.G.B., Liu, H.: Where are they going? Predicting human behaviors in crowded scenes. ACM Trans. Multimedia Comput. **17**, 1–19 (2021)
11. Zhang, B., Wang, N., Zhao, Z., Abraham, A., Liu, H.: Crowd counting based on attention-guided multi-scale fusion networks. Neurocomputing **451**, 12–24 (2021)
12. Zernetsch, S., Kohnen, S., Goldhammer, M., Doll, K., Sick, B.: Trajectory prediction of cyclists using a physical model and an artificial neural network. In: 2016 IEEE Intelligent Vehicles Symposium, pp. 833–838. IEEE (2016)
13. Kooij, J.F.P., Flohr, F., Pool, E.A.I., Gavrila, D.M.: Context-based path prediction for targets with switching dynamics. Int. J. Comput. Vis. **127**(3), 239–262 (2019). https://doi.org/10.1007/s11263-018-1104-4
14. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social LSTM: human trajectory prediction in crowded spaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 961–971 (2016)
15. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social GAN: socially acceptable trajectories with generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2255–2264 (2018)
16. Mohamed, A., Qian, K., Elhoseiny, M., Claudel, C.: Social-STGCNN: a social spatio-temporal graph convolutional neural network for human trajectory prediction. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition, pp. 14424–14432 (2020)
17. Shen, M., Habibi, G., How, J.P.: Transferable pedestrian motion prediction models at intersections. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 4547–4553, IEEE (2018)
18. Ho, J., Ermon, S.: Generative adversarial imitation learning. Adv. Neural Inf. Process. Syst. **29**, 4565–4573 (2016)
19. Rhinehart, N., Kitani, K.M., Vernaza, P.: R2P2: a reparameterized pushforward policy for diverse, precise generative path forecasting. In: Proceedings of the European Conference on Computer Vision, pp. 772–788 (2018)
20. Xie, G., Shangguan, A., Fei, R., Ji, W., Ma, W., Hei, X.: Motion trajectory prediction based on a CNN-LSTM sequential model. Sci. China Inf. Sci. **63**(11), 212207 (2020). https://doi.org/10.1007/s11432-019-2761-y
21. Zhao, D., Jean, O.: Noticing motion patterns: a temporal CNN with a novel convolution operator for human trajectory prediction. IEEE Robot. Autom. Lett. **6**(2), 628–634 (2020)
22. Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Rezatofighi, H., Savarese, S.: SoPhie: an attentive GAN for predicting paths compliant to social and physical constraints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1349–1358 (2019)