IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 August 2022, accepted 16 August 2022, date of publication 18 August 2022, date of current version 29 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3200066

METHODS

Multimodal Pedestrian Trajectory Prediction Based on Relative Interactive Spatial-Temporal Graph

DUAN ZHAO[1,2], TAO LI[1,2], XIANGYU ZOU[1,2], YAOYI HE[3], LICHANG ZHAO[3], HUI CHEN[3], AND MINMIN ZHUO[3]

[1]School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, Jiangsu 221008, China
[2]The National Joint Engineering Laboratory of Internet Applied Technology of Mines, Xuzhou, Jiangsu 221008, China
[3]Tiandi (Changzhou) Automation Company Ltd., Changzhou, Jiangsu 213000, China

Corresponding author: Yaoyi He (hyy@cari.com.cn)

This work was supported by the Science and Technology Innovation and Entrepreneurship Fund Project of Tiandi Technology Company Ltd., under Grant 2019-TD-ZD007.

ABSTRACT Predicting and understanding pedestrian intentions is crucial for autonomous vehicles and mobile robots to navigate in a crowd. However, the movement of pedestrian is random. Pedestrian trajectory modeling needs to consider not only the past movement of pedestrians, the interaction between different pedestrians, the constraints of static obstacles in the scene, but also multi-modal of the human trajectory, which brings challenges to pedestrian trajectory prediction. Most of the existing trajectory prediction methods only consider the interaction between pedestrians in the scene, ignoring the static obstacles in the scene can also have impacts on the trajectory of pedestrian. In this paper, a scalable relative interactive spatial-temporal graph generation adversarial network architecture (RISTG-GAN) is proposed to generate a reasonable multi-modal prediction trajectory by considering the interaction effects of all agents in the scene. Our method extends recent work on trajectory prediction. First, LSTM nodes are flexibly used to model the spatial-temporal graph of human-environment interactions, and the spatial-temporal graph is converted into feed-forward differentiable feature coding, and the time attention module is proposed to capture the trajectory information in time domain and learn the time dependence in long time range. Then, we capture the relative importance of the interaction of all agents in the scene on the pedestrian trajectory through the improved relative scaled dot product attention and use the generative adversarial network architecture for training to generate reasonable pedestrian future trajectory distribution. Experiments on five commonly used real public datasets show that RISTG-GAN is better than previous work in terms of reasoning speed, accuracy and the rationality of trajectory prediction.

INDEX TERMS Pedestrian trajectory prediction, spatial-temporal graph, time attention, relative scaled dot product attention, generative adversarial network.

I. INTRODUCTION

With the development of society, autonomous navigation platforms like autonomous vehicles and social robots are growing, it is critical that autonomous vehicles and social robots must be able to predict the movements of pedestrians to prevent collisions with them [1], [2], [3], [4]. So, modeling the behaviors of pedestrians is an essential step for autonomous platforms application research, such as video autonomous monitoring platform detection suspicious trajectory [5], [6], [7], socially-aware robots for visual navigation [8], [9], and self-driving platforms safety decisions [10], [11], [12]. Pedestrian trajectory prediction is defined as the prediction of pedestrian movement trajectory for some time to come based

The associate editor coordinating the review of this manuscript and approving it for publication was Jjun Cheng.

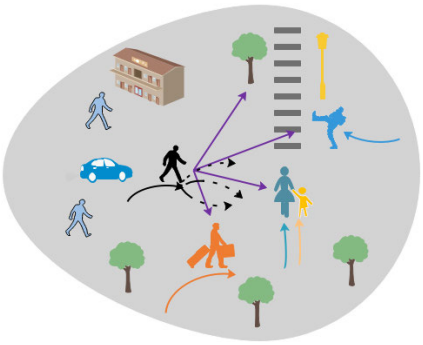


FIGURE 1. An example of pedestrian multi-modal trajectory in a crowded scene. In the scene, pedestrians will pay more attention to the person or object in front of them and pay less attention to the situation behind them. When making navigational decisions, pedestrians judge and analyze the importance of dynamic participants and static obstacles in the scene (such as trees, stationary vehicles, and streetlights) for future trajectory interactions. It should be noted that the scene changes dynamically, and so does the trajectory of pedestrians at each moment, with randomness. so, we should not only consider the spatio-temporal structure of human interaction with the current scene, visual attention, but also the random multi-modal nature of human walking.

on the past trajectory of pedestrian, accurate trajectory prediction can help autonomous driving and social robots navigate better.

Predicting the trajectory of pedestrian in a crowded scene is a challenging task. It is necessary to consider the spatial-temporal structure of human interaction with the current scene, visual attention [13] that human can quickly analyze the environment, and the random multi-modal [14] nature of human walking. In the process of walking, pedestrians can walk according to the intentions of surrounding neighbors and the positions of static obstacles to adjust their own trajectory to avoid collision, and with the passage of time, the scene of the pedestrians constantly moving, obstacles types and relative positions are constantly changing, pedestrians must also constantly adjust their own trajectory, so the interaction between human and dynamic environment has spatial-temporal structure [15]. In addition, humans are born with the ability to analyze and judge, people or objects that are nearby are more likely to attract the target pedestrian's attention than distant people or objects, or the target pedestrian pays more attention to the people in front of him than to the people behind him, in what's known as the ''attentional mechanism.'' In view of this situation, Laurent Itti *et al.* [13] proposed a visual attention model that can explain this situation. In addition, according to the actual situation in real life, the movement trajectory of pedestrians will be more random and uncertain due to the influence of other pedestrians or obstacles in the scene, as shown in Fig. 1. Compared with the deterministic trajectory prediction proposed before, the multi-modal trajectory prediction output is more reasonable.

However, most of the existing trajectory prediction methods do not consider the above three aspects simultaneously. Early work on pedestrian trajectory prediction mainly focused on model-driven methods. Helbing *et al.* [16] pioneered the social force model, which predictes pedestrian behavior according to attraction and repulsion. Morris *et al.* [17] proposed the Hidden Markov Model to predict pedestrian trajectory with spatial-temporal probabilities. However, these models have poor adaptability and are too sensitive to parameters, so they can not solve the problem of pedestrian trajectory prediction in crowded scenes. In recent years, data-driven method has become a popular research method for researchers. This method mainly regards pedestrian trajectory prediction as a time series generation task. Some recent works have used recursive neural networks (RNNS) to solve this problem. Alahi *et al.* [18] proposed the Social-LSTM model and innovatively used the social pooling layer module to divide the space where pedestrians are with rectangular grid units, so that capture the interactive information of adjacent pedestrians. Bisagno *et al.* [19] proposed the Group-LSTM model, which is an improved method of Social-LSTM. This model uses motion consistency to gather trajectories with similar movement trends and to group pedestrians. However, the above methods only consider the simple interaction between pedestrians and fail to capture the spatial-temporal interaction information between human and the current scene.

In view of the spatial-temporal interaction between human and the environment, [20], [21], [22] proposed a modeling method based on spatial-temporal graph (STG), through which the temporal and spatial connection between the target subjects can be clearly modeled. Mohamed *et al.* [23] proposed social spatial-temporal graph convolutional neural network (social-STGCNN), which models the interaction between pedestrians as spatial-temporal graph to replace the aggregation method, but they did not consider scene modeling. Sirin Haddad *et al.* [24] proposed a long and short term-memory (LSTM) network based on spatial-temporal graph, the interaction of all agents in the scenario was considered. Although the above methods model the spatial-temporal interaction between human and environment through spatial-temporal graph, it is the deterministic trajectory prediction output. Considering the randomness and uncertainty of pedestrian movement, the multi-modal trajectories prediction output is realistic and reasonable.

Since generative adversarial networks (GANs) [25] have achieved remarkable results in probability calculation and behavioral reasoning, researchers begin to turn their attention to GAN. Aglim Gupta *et al.* [26] proposed a pioneering social-GAN model and introduced GAN into the pedestrian trajectory prediction task. This model combines sequence prediction with generative adversarial network to generate diversified trajectories, and generated reasonable trajectory prediction through repeated adversarial training. However, the model do not consider global scenario information. Stuart Eiffert *et al.* [27] proposed a probabilistic crowd GAN (PCGAN) trajectory prediction method, which combines the recursive neural network and the mixed density network. This method not only consideres the interaction effects between people and people, but also the interaction effects between people and vehicles, but do not model the time information, and the training process is very complicated.

In order to solve the limitation of the above methods, we extend our previous work Zou *et al.* [28], introducing a scalable relative interactive spatial-temporal graph generation adversarial network architecture (RISTG-GAN), which comprehensively considers the interaction effects of all agents in the scene to generate reasonable multi-modal prediction trajectory. First, LSTM nodes are flexibly used to model the spatial-temporal graph of human-environment interactions, and the spatial-temporal graph is converted into feed-forward differentiable feature coding. Then, we introduce the time attention module to assign different weights to the past trajectory sequence of pedestrians, extract important information at different moments and weaken the speed deviation of different pedestrians. Finally, improved relative scaled dot product attention is used to capture the relative importance of various interactions in the scene on pedestrian trajectory and use recurrent sequence modeling and generative adversarial network for joint training to generate reasonable future trajectory prediction output. The main contributions of this paper are as follows:

1) This paper proposes a scalable RISTG-GAN architecture, and the number of nodes can change dynamically according to different scenes. The framework models all interactions in the scene and uses recurrent sequence modeling and generation adversarial network architecture to train together to generate multi-modal pedestrian trajectory prediction, which conforms to the characteristics of randomness and uncertainty of pedestrian walking in the real scene.
2) In the feature coding stage, a time attention module is introduced to assign different weights to the past trajectory sequence of pedestrians, extract important information at different moments, align the pedestrians in the space, and weaken the speed deviation of different pedestrians.
3) In the interaction stage, an improved scaled dot product attention is introduced to capture the relative importance of the impacts of all interactions on the pedestrian trajectory in the scene, which is more in line with the innate characteristics of human beings to screen information.

The rest of this paper is arranged as follows. In Section II, we analyze recent work on pedestrian trajectory prediction. In Section III, we explain the principle of the RISTG-GAN pedestrian trajectory prediction model in detail. In Section IV, we do comparative experiments with other models on the open data sets and analyze the experimental results. In Section V, we summarize the work of this paper.

II. RELATED WORK

The focus of our work is to predict the trajectory of pedestrians. In the past decades, many researchers have carried out research on pedestrian trajectory prediction and put forward their own methods. Previous work has focused on modeling with hand-made feature functions [16], [29], [30], [31]. However, with the rapid development of deep learning, data-driven methods based on deep learning have recently made great progress in trajectory prediction. In this section, we focus on RNN-related sequence prediction, attention mechanism and GAN model related to our work.

A. RECURRENT NEURAL NETWORKS FOR SEQUENCE PREDICTION

Pedestrian trajectory prediction is defined as predicting the future movement trajectory of pedestrian according to the past trajectory, which is a typical sequence generation problem. In recent years, recurrent neural network (RNN) has achieved great success in the task of sequence prediction. As a variant of RNN, long and short-term memory network (LSTM) [32] can learn long-term dependencies. LSTM has designed three ''gate'' structures to control the cell state, namely forgetting gate, input gate and output gate. The function of the forgetting gate is to decide what information to discard from the cell state, thus solving the problem of large computational data and noise. Input gate is the selective memory stage, its function is to selectively ''remember'' the input, important information is recorded, otherwise less memory. The function of the output gate is to decide what information to output from the cell state. Thanks to the excellent application of LSTM in machine translation [33] and speech recognition [34], researchers begin to widely apply LSTM to the prediction of pedestrian trajectory. Alahi *et al.* [18] first proposed the Social-LSTM model, the space where pedestrians are was divided by rectangular grid units and used the social pooling layer to capture the interactions between pedestrians, allowing neighboring pedestrians to share the hidden state. Huynh Manh *et al.* [35] proposed scene-LSTM, which combines the scene information and the historical trajectory of pedestrian to predict the future trajectory of pedestrian in static crowded scenes. Xue *et al.* [36] proposed SS-LSTM, which uses three different LSTM networks to capture pedestrian, social and scene size information respectively to improve the ability to predict pedestrian trajectory.

Although the above methods improve the pedestrian trajectory prediction ability to different degrees, they do not consider different pedestrians or objects have different degrees of impact on the target pedestrian.

B. ATTENTION MECHANISMS ARE USED FOR TRAJECTORY PREDICTION

Humans are born with the ability of analysis and judgment. When walking in a crowded scene, pedestrians will pay more attention to the nearby people and obstacles in front of them compared with the pedestrians or obstacles behind them or in the distance. This is because humans use limited visual attention to quickly screen out useful information from the scene, so the attention mechanism is proposed. Thanks to the successful application of attention mechanism in natural language processing [37], some researchers have introduced attention mechanism into the field of pedestrian trajectory prediction, capturing the relative importance of neighbors and obstacles around pedestrians in the scene. Vemula *et al.* [38]

proposed a social attention mechanism that can capture the relative importance of the current pedestrian navigation of other pedestrians in the scene. Fernando *et al.* [39] proposed a combination method of soft attention and hard attention. Soft attention was used to evaluate the significance of interaction in the scene area, and hard attention was used to assign different weights to pedestrians at different distances. Velickovic *et al.* [40] proposed a graph attention mechanism, in which stacked nodes can pay attention to the layer of their neighborhood characteristics and assign different weights to different nodes in the neighborhood. Sirin Haddad *et al.* [24] proposed the spatial-temporal attention mechanism, which is a variant of the multi-head method. It retains the global interaction information of all pedestrians in the scene and the local interaction information of static objects in the way of accumulation and average. Stuart Eiffert *et al.* [27] used the Graph Vehicle-Pedestrian Attention Network (GVAT) to focus on a much wider range of problems: pedestrians and vehicles. The network models social interactions and allows input of shared vehicle characteristics. These methods indicate that the introduction of attention mechanism can indeed improve the accuracy of pedestrian trajectory prediction. In our work, we capture pedestrian trajectory information in the time domain ang learn time dependence over long time ranges by introducing temporal attention. Recently, Transformer Networks have made great strides in Natural Language Processing [41], [42], we borrowed this method, introducing relative scaled dot product attention to capture the relative importance of various interactions in a global scene affecting pedestrian trajectories.

C. GENERATING ADVERSARIAL NETWORKS (GANS)

The above methods are the only deterministic trajectory prediction output. However, in real life, the trajectory of pedestrians shows more randomness and uncertainty due to the influence of other pedestrians or obstacles in the scene, and the multi-mode trajectory prediction output is more consistent with the real situation. Initially generative adversarial networks (GANs) [25] were used in probability calculation and behavioral reasoning. Agrim Gupta *et al.* [26] introduced GAN into the pedestrian trajectory prediction task for the first time and proposed a social-GAN model. The generator is composed of an LSTM based encoder-decoder with a social pool layer that simulates the relationship between each pedestrian. The discriminator distinguishes whether the generated trajectory is real (ground real) or false (generated) and generates reasonable trajectory prediction through repeated adversarial training. Amir Sadeghian *et al.* [43] extended this idea and improved the model by adding physical and social attention mechanism. The improved model can extract the most important trajectory information from the neighbors and assign different soft attention weights to the static environment. Vineet Kosaraju *et al.* [44] proposed the Social-BiGAT model and introduced a generative adversarial network based on graph to better simulate the social interaction of pedestrians in the scene through flexible graph structure to generate reasonable multi-modal trajectory prediction.

III. PROBLEM REPRESENTATION AND MODEL

In this section, we first define the pedestrian trajectory prediction problem. Next, introducing the RISTG-GAN framework, and then describe the working principle of the spatial-temporal graph feature coding based on relative interaction. Finally, this paper illustrates the process of using recurrent sequence modeling and generation adversarial network to train together to output reasonable trajectory prediction.

A. PROBLEM DEFINITION

In this paper, we aim at the prediction of pedestrian trajectory (x and y coordinates on a 2D map) in a fixed scene, and comprehensively consider the previous movement of pedestrians and the position of fixed obstacles in the scene (including stationary vehicles, lamp posts etc). At every moment, pedestrians regard the positions of other pedestrians and obstacles around them as a static "map". With the change of time, these static maps become a dynamic map with temporal sequence information. Therefore, the interaction between people and the environment has a spatial-temporal structure. The observable historical trajectory of pedestrian i is defined as: $X_i=\{(x_i^t, y_i^t) \mid t = 1, \ldots, t_{obs}\}$, fixed obstacles oi observable historical position is defined as: $Z_{oi}=\{(x_{oi}^t, y_{oi}^t) \mid t = 1, \ldots, t_{obs}\}$, the real future trajectory of pedestrian i is defined as: $Y_i=\{(x_i^t, y_i^t) \mid t = t_{obs+1}, \ldots, t_{pred}\}$. Similarly, the predicted future pedestrian trajectory is defined as: $\hat{Y}_i=\{(\hat{x}_i^t, \hat{y}_i^t) \mid t = t_{obs+1}, \ldots, t_{pred}\}$.

B. OVERALL MODEL

This paper proposes a new pedestrian trajectory prediction method, RISTG-GAN, which considers the historical trajectory, state, interaction of surrounding pedestrians and fixed obstacles of each pedestrian in the scene comprehensively that can accurately predict the pedestrian trajectory. The overall architecture is shown in Fig. 2. The architecture can be divided into three modules, which are the feature encoder module, generator/decoder module and discriminator module. The feature encoder module includes the time attention module and the relative scaled dot product attention module. First, the interaction model of dynamic participants and fixed obstacles in the scene is established by using the spatial-temporal graph, and LSTM is used to extract nodes feature coding from the historical trajectory information of pedestrians. Next, the extracted feature coding is input into the time attention module, and different weights are assigned to it in each time step to get the time feature coding. Finally, we improve the scaled dot product attention proposed in reference [45] and propose the relative scaled dot product attention. The historical trajectory information of pedestrian, the location information of fixed obstacle and the time feature coding are input into the relative scaled dot product attention

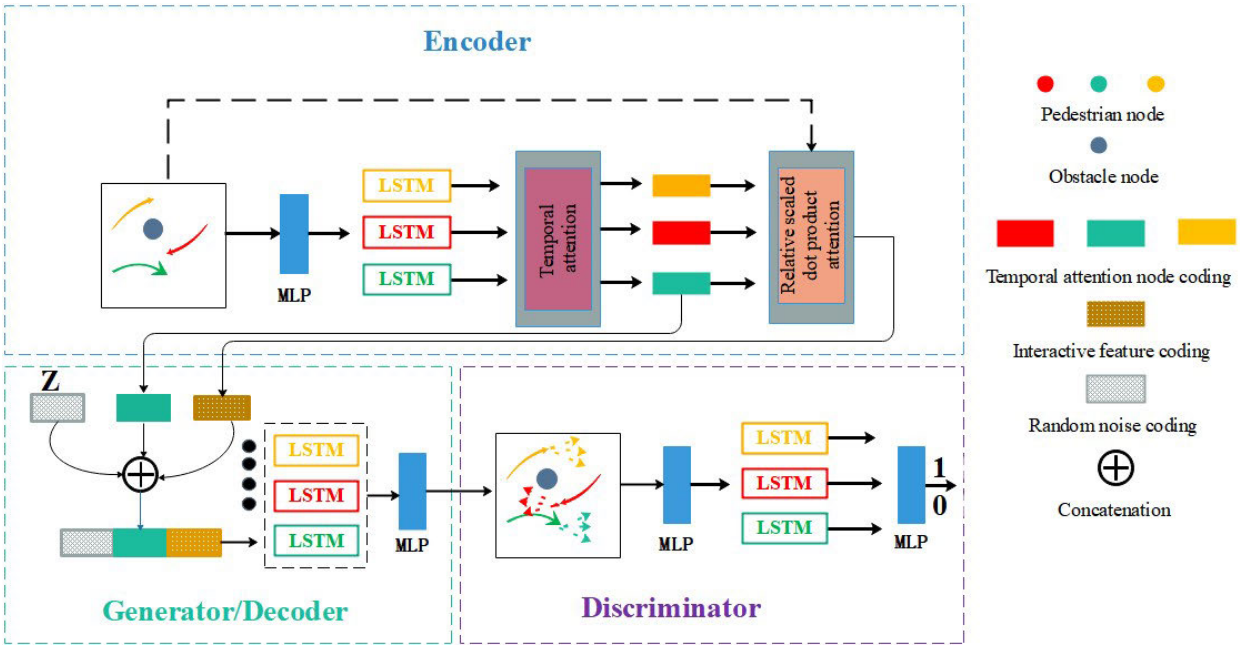


FIGURE 2. Our proposed relative interactive spatial-temporal graph network architecture.The framework consists of three main modules:feature encoder module, generator/decoder module and discriminator module.

module to capture the relative interactive feature encoding of the impact of the global scene on the pedestrian trajectory.

In the generator/decoder module, we integrate random noise, time feature coding and relative interactive feature coding as the input of the generator/decoder module. Based on these features, the generator generates a distribution of diversity trajectories that conform to social rules. In the discriminator module, the discriminator is repeatedly trained to distinguish whether the generated trajectory distribution is real (ground true) or false (generated), and when the discriminator cannot clearly distinguish between the generated trajectory true and false, the output is reasonable.

C. SPATIAL-TEMPORAL GRAPH ARCHITECTURE

In this paper, we describe the dynamic spatial-temporal structure of the interaction between pedestrians and the environment in the scene by using spatial-temporal graph. We express the spatial-temporal graph as: $G = (v, \varepsilon_S, \varepsilon_T)$, Where v is the instance nodes set, ε_T is a set of time edges, ε_S is a set of spatial edges, and its abstract network architecture is shown in Fig. 3. In the paper, instance nodes include pedestrian node P and fixed obstacle node O, the nodes is variable. The spatial edge connects all instance nodes, while the time edge connects adjacent time steps to the same pedestrian node. It is worth noting that the obstacle nodes do not need to be connected at adjacent time steps, because the position of the obstacle does not change with time.

In this paper, we introduce the time attention module to capture the temporal edge information of pedestrian trajectory and use the improved relative scaled dot product attention to capture the spatial edge information. We will introduce the two aspects respectively below.

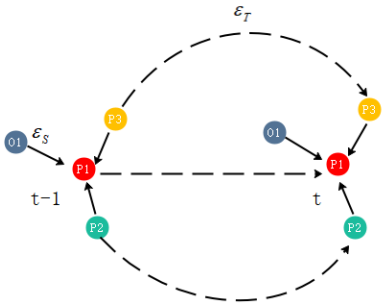


FIGURE 3. Structure of spatial-temporal interaction information for pedestrians with adjacent time steps. The spatial relationship between pedestrians and obstacles is represented by a black solid arrow, ε_T represents the spatial edge, the black dotted line represents the time edges that connects the same pedestrian node on adjacent time steps, ε_S represents the time edge.

1) TIME ATTENTION MOUDLE

In the pedestrian trajectory prediction task, the position of the pedestrian changes dynamically with time, so it is necessary to capture the trajectory information in the time domain. By introducing the time attention module, we extract the trajectory information in the time domain and assign different weights to it. Taking pedestrian i as an example, we first use multi-layer perceptron (MLP) to embed coordinate position of pedestrian i to obtain fixed length vector e_{pi}^t and LSTM unit takes this embedded vector as input to obtain pedestrian node feature code h_i^t.

$$e_{pi}^t = \phi(x_i^t, y_i^t; W_p) \tag{1}$$

$$h_i^t = LSTM(h_i^{t-1}, e_{pi}^t; W_{temporal}^p) \tag{2}$$

where $\phi(\cdot)$ is a nonlinear embedding function, W_p is the embedding weight, $W_{temporal}^p$ is the weight of the temporal-edge LSTM cell.

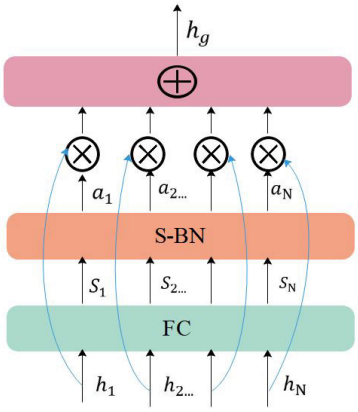


FIGURE 4. Network architecture of time attention module.

We take the pedestrian node feature code h_i^t obtained above as the input of the time attention module. Fig. 4 shows the network architecture of our time attention module. Where N represents the number of people in the scene and node feature $h_i (i = 1, 2, \ldots .N)$ is input to the FC layer to obtain score S_i.

$$S_i = FC(h_i) = tanh(w_a^T h_i + b_a) \tag{3}$$

where FC is the fully connected network, S_i is the calculation of the score of h_i, w_a and b_a are the network parameters, and $tanh()$ is the activation function. Next, S_i is taken as the input of $S - BN$ layer, and the attention weight a_i of h_i is obtained.

$$a_i = softmax(BN(S_i)) \tag{4}$$

where BN is the Batch Normalization function and S is the $softmax()$ function. Finally, the time feature coding vector $\hat{h}_i$ with time information is obtained by multiplying the respective node feature coding h_i and its corresponding attention weight a_i and summation.

$$\hat{h}_i = \sum_{i=1}^{N} (a_i h_i) \tag{5}$$

We capture the time edge information of pedestrian trajectory through the time attention module, which improves the accuracy and robustness of the model.

2) RELATIVE SCALED DOT PRODUCT ATTENTION MODULE

In the pedestrian trajectory prediction task, the trajectory of the target pedestrian is not only affected by the surrounding pedestrians, but also by the fixed obstacles in the scene, so we introduce the relative scaled dot product attention module to capture the spatial information of all instance nodes. It considers not only the relative position of the target pedestrian and its neighbors in current and historical moments, but also the relative position with the fixed obstacles, and assigns different weights. First, we calculate the relative distance O_{ij}^t between the pedestrian and the fixed obstacle node.

$$O_{ij}^t = \begin{cases} (x_i^t - x_{oi}^t, y_i^t - y_{oi}^t) & \text{Obstacles exist} \\ (0, 0) & \text{Obstacles do not exist} \end{cases} \tag{6}$$

Next, the fixed length vector r_{ij}^t is obtained by embedding the relative distance from pedestrian i to adjacent pedestrian j and to the obstacle through multi-layer perceptron, and then the vector r_{ij}^t is used as the input of LSTM unit to obtain the relative feature code h_r^t, When the obstacle nodes exist, the relative feature code h_r^t contains the context information of the scene. When obstacle nodes do not exist, the relative feature coding is reduced to contain only social interaction information.

$$r_{ij}^t = \phi(x_i^t - x_j^t, y_i^t - y_j^t, O_{ij}^t; W_r) \tag{7}$$

$$h_r^t = LSTM(h_r^{t-1}, r_{ij}^t; W_{spatial}^r) \tag{8}$$

where, W_r is the embedded weight, and $W_{spatial}^r$ is the weight of the spatial edge-LSTM cell, which is shared among all instance nodes. Then, we use the scaled dot product attention mechanism proposed in literature [45] to assign influence weight to all instance nodes in the scene. Finally, the influence weight is multiplied by the time feature coding vector $\hat{h}_i$ to obtain the relative interaction feature coding I_i^t (the yellow grid square in Fig. 2).

$$W_R^t = softmax(\frac{1}{\sqrt{d_e}} Dot(W_2 h_r^t, W_1 \hat{h}_i^t)) \tag{9}$$

$$I_i^t = W_R^t \cdot h_r^t \tag{10}$$

where W_1 and W_2 are weights used for linear scaling and projection of hidden states onto the d_e dimension vector, $Dot(\cdot)$ is the dot product, $softmax()$ is the activation function, $\frac{1}{\sqrt{d_e}}$ scaling factor. So far, the process of encoding relative interactive features based on spatial-temporal graph has been completed.

D. GENERATOR

As mentioned in the introduction, pedestrian trajectories in crowded scene are stochastic and uncertain, so it is reasonable to use multimodal trajectory prediction output. Generative adversarial network is used for training. For the generator module (G), as shown in Fig. 2, we use the decoder based on LSTM unit for eigenvector decoding and trajectory generation. First, we introduce the standard normally distributed noise z (the gray square in Fig. 2). Next, we connect the time feature coding vector $\hat{h}_i$, the relative interactive feature coding I_i^t and the noise vector z as the input of the decoder LSTM unit to obtain the mixed feature coding vector h_{gi}^t. Then, h_{gi}^t is converted to spatial coordinates through a multi-layer perceptron.

$$h_{gi}^t = LSTM(h_{gi}^{t-1}, [z, \hat{h}_i^t, I_i^t]; W_g) \tag{11}$$

$$(\hat{x}_i^{t+1}, \hat{y}_i^{t+1}) = MLP(h_{gi}^t; W_{ge}) \tag{12}$$

where z is the noise vector satisfying the standard normal distribution, $MLP(\cdot)$ is the multi-layer perceptron, W_g and W_{ge} are the embedding weights.

E. DISCRIMINATOR

For the discriminator module (D), as shown in Fig. 2. Based on the observation of all the historical trajectories of

pedestrians, the discriminator will evaluate the real future trajectories of pedestrians Y_i and the predicted future trajectories $\hat{Y}_i$. We use *MLP* in the last hidden state of the encoder to get the classification score L_{disi}.

$$e^t_{disi} = MLP([T^t_i]; W_{e1}) \tag{13}$$
$$h^t_{disi} = LSTM(h^{t-1}_{disi}, e^t_{disi}; W_{e2}) \tag{14}$$
$$L_{disi} = MLP(h_{disi}; W_{e3}) \tag{15}$$

where T^t_i is each coordinate from $[X^t_i, \hat{Y}^t_i]$ or $[X^t_i, Y^t_i]$ at time t, h_{disi} is the integration h^t_{disi}, L_{disi} is the result of the classification (true/false), When $L_{disi} = 0$, it means that the output trajectory is false; when $L_{disi} = 1$, it means that the output trajectory is real, W_{e1}, W_{e2}, and W_{e3} are embedding weights, respectively.

F. LOSS

We defined the training goals of RISTG-GAN as follows:

$$V = arg \min_G \max_D L_{GAN}(G, D) + \lambda L_{L2}(G) \tag{16}$$

where λ is the weighting coefficient, and the adversarial loss $L_{GAN}(G, D)$ and L2 loss $L_{L2}(G)$ are defined as follows:

$$L_{GAN}(G, D) = E_{i \in RISTG}[logD(Y_i)] + E_{i \in RISTG}[log(1 - D(\hat{Y}_i))] \tag{17}$$
$$L_{L2}(G) = E_{i \in RISTG}[\left\| Y_i - \hat{Y}_i \right\|_2] \tag{18}$$

G. IMPLEMENTATION DETAILS

In our proposed model, the encoder and decoder are constructed based on LSTM units. The hidden state sizes of the encoder and decoder are 16 and 32 respectively, and the input coordinates are embedded into 16-dimensional vectors. The ADAM optimizer [46] is used to train the generator and discriminator models, the initial learning rate is set to 0.001, and the number of training are set to 200 rounds.

IV. EXPERIMENTS

In this section, we introduce the two data sets used in the experiment and the measurement criteria, showing the experimental results of our method, comparing its performance with the most advanced method, and showing the quantitative analysis and qualitative results.

A. DATASETS AND METRICS

We evaluate the model performance on two common pedestrian trajectory datasets: ETH [47] and UCY [48]. The ETH dataset contains two subsets named Eth-univ and Eth-hotel, while the UCY dataset contains three subsets named UCY-zara1, UCY-zara2 and UCY-univ. These five real scenes contain the interactions between most people and the environment in the real world, such as turning at the intersection, following the crowd, avoiding the obstacles on the road, and intersecting with each other. Same work as in [26], [36], and [49], we use data within 8 seconds to evaluate the model and record a time step of 0.4 seconds. Among them, the first 3.2 seconds (8 time steps) are training data, and the last 4.8 seconds (12 time steps) are test data. In the evaluation of the model, we use two benchmark metrics: the mean displacement error (ADE) and the final displacement error (FDE). The mean displacement error (ADE) is defined as the average L2 distance between the ground reality and our prediction over all predicted time steps.

$$ADE = \frac{\sum_{i=1}^{N} \sum_{t=t_{obs}+1}^{t_{pred}} \left\| Y^t_i - \hat{Y}^t_i \right\|_2}{N * (t_{pred} - t_{obs} - 1)} \tag{19}$$

The final displacement error (FDE) is defined as the mean distance between the predicted final destination and the true final destination.

$$FDE = \frac{\sum_{i=1}^{N} \left\| Y_i^{t_{pred}} - \hat{Y}_i^{t_{pred}} \right\|_2}{N} \tag{20}$$

where, $\hat{Y}^t_i$ and Y^t_i are the predicted position and real position of pedestrian i at time t respectively, and N is the number of pedestrians in the scene. The smaller ADE and FDE values are, the more accurate trajectory prediction is.

In this paper, in order to test the validity of the model, we choose five models for comparison, including LSTM [32], S-LSTM [18], SS-LSTM [36], S-GAN [26], and Sophie [43]. In addition, we have also performed ablation research on the proposed RISTG-GAN. In the RISTG-GAN framework, we model the complex interaction between people and the environment through the spatial-temporal graph, the instance nodes in the scene are divided into pedestrian nodes and obstacle nodes. The method that only considers pedestrian nodes is called RISTG-GAN-1, and the method that considers all agents is called RISTG-GAN-2. In Table 1, we describe the modeling direction of the seven models respectively.

B. QUANTITATIVE EVALUATION

In Table 2, our proposed model is compared with other five existing typical models on five publicly available datasets. We use data within 8 seconds to evaluate the model, taking the first 3.2 seconds of each trajectory as the training value and predicting the next 4.8 seconds of trajectory. Through comparison, it can be found that the LSTM model has the worst performance, because the model only considers the historical trajectory of pedestrian. The performance of S-LSTM model is better than that of the simple LSTM model, because the model proposes to use the social-pooling layer to capture the interaction information between local pedestrians. Compared with S-LSTM model, SS-LSTM model not only considers the interaction of all pedestrians in the scene, but also uses the context information of the scene to predict the pedestrian trajectory. The average values of ADE and FDE in the five data sets decrease by 18%and 17%, respectively, which further proves the importance of considering the context information of the scene for prediction.

Compared with the above LSTM-based prediction model, the prediction error of GAN-based prediction model is

TABLE 1. Analysis of seven benchmark evaluation models.

Model	Innovation	Modeling Perspectives
LSTM	Data driven nonlinear trajectory prediction method	Individual
S-LSTM	A social-pooling layer is proposed to capture local social interactions	Individual + interaction
SS-LSTM	Considering the impact of scene information on pedestrian trajectory	Individual + interaction + scene information
S-GAN	The multi-modality property of pedestrian trajectory is considered, and a new pooling layer is introduced;	Individual + interaction multi-modality
SoPhie	Combining physical attention mechanism and social attention mechanism, the scene information is considered;	Individual + interaction Scene information + multi-modality
RISTG-GAN-1	1.The interactive information is modeled through the spatial-temporal graphs, and the instance nodes can be flexibly changed according to the input;	Individual + interaction social attention + multi-modality
RISTG-GAN-2	2.Time attention module is used to capture the information of pedestrian trajectory; 3.Dot product attention is used to capture the relative importance of various impacts.	Individual + scene context scene attention + multi-modality

TABLE 2. Quantitative results for all models on five open datasets, with ADE and FDE values separated by slashes.

	Baselines					RISTG-GAN(Ours)	
Dataset	**LSTM**	**S-LSTM**	**SS-LSTM**	**S-GAN**	**SoPhie**	**RISTG-GAN-1**	**RISTG-GAN-2**
ETH-univ	1.05/2.39	1.09/2.35	0.92/1.92	0.81/1.52	0.70/1.43	0.65/1.35	**0.58/1.29**
ETH-hotel	0.82/1.92	0.79/1.76	0.67/1.5	0.72/1.61	0.76/1.67	0.62/1.41	**0.49/1.12**
UCY-zara1	0.43/0.91	0.47/1.00	0.38/0.75	0.34/0.69	0.54/1.24	0.34/0.69	0.36/0.72
UCY-zara2	0.54/1.15	0.56/1.17	0.41/0.8	0.42/0.84	0.30/0.63	**0.28/0.58**	**0.26/0.57**
UCY-univ	0.83/1.75	0.67/1.40	0.57/1.38	0.60/1.26	0.38/0.78	0.36/**0.79**	**0.35**/0.82
AVG	0.75/1.62	0.72/1.54	0.59/1.27	0.58/1.18	0.54/1.15	0.45/0.96	**0.41/0.90**

TABLE 3. Comparison of reasoning speed of seven models.

	LSTM	S-LSTM	SS-LSTM	S-GAN	RISTG-GAN-1	RISTG-GAN-2
Inference time(s)	**0.045**	2.542	2.851	0.126	0.132	0.135
Speed-up	**76.1x**	1.2x	1x	27.2x	23.8x	21.4x

smaller. S-GAN model is the first to introduce the generative adversarial network into the pedestrian trajectory prediction task, the model considers the multi-modality property of pedestrian trajectory in crowded scene and proposes a new pooling layer, so its performance is better than that of SS-LSTM. Based on the S-GAN model, Sophie model takes the scene information into account and improves the prediction performance of the model by introducing the physical attention mechanism and the social attention mechanism, especially on the UCY-zrar2 dataset. Compared with Sophie, the RISTG-GAN-1 models complex scene by using spatial-temporal graph and captures the relative importance of crowd interaction to pedestrian trajectory by using relative scaled dot product attention. The results show that trajectory prediction errors are further reduced. Based on the RISTG-GAN-1 model, the RISTG-GAN-2 model considers the positions of fixed obstacles in the scene, because the real scene contains not only moving pedestrians, but also stationary obstacles (such as lamp posts and stationary vehicles), so the prediction performance is further improved. By observing table 2, it is found that although the RISTG-GAN-2 model considers fixed obstacle nodes, there is no significant difference between the RISTG-GAN-1 model and the RISTG-GAN-1 model in the evaluation performance of the three data sets UCY-zara1/zara2/univ, which may be because the position of obstacles in the scene contained in these data sets has little influence on the walking of pedestrians. In the data sets ETH-univ and ETH-hotel, the RISTG-GAN-2 model performs better, because there are many obstacles in the scene in these two data sets, which pedestrians need to avoid. Compared with the SoPhie model, the mean values of ADE and FDE of the RISTG-GAN-2 model on five data sets are reduced by 24% and 21.7%, respectively.

For autonomous vehicles and social robots, it is crucial to accurately and quickly reason out the next trajectory of pedestrian to avoid collisions in crowded scene. The faster the reasoning speed is, the further guarantee of pedestrian safety can be obtained. Therefore, we also compare the reasoning speed of the models. In Table 3, we record the speed of reasoning for each model. Because the LSTM model only considers the historical trajectory of pedestrian, the amount of reasoning tasks is small, so the reasoning speed is the fastest, but the accuracy is too low. Both S-LSTM model and SS-LSTM model are improved base on LSTM model. The space where pedestrians are located is divided by grid cells and the interaction is calculated. The calculation efficiency is low and the reasoning speed is the slowest. S-GAN model introduces the generative adversarial network (GAN), which considers the multimodal property of pedestrian trajectory, and not only has high accuracy, but also the reasoning speed is fast, because adversarial training can significantly improve the memory utilization rate. Based on this, our model also combined with GAN to model the scene through flexible spatial-temporal graph. Compared with the SS-LSTM model, the reasoning speed of RISTG-GAN-1 and RISTG-GAN-2 is increased by 18.83 times and 16.83 times respectively.

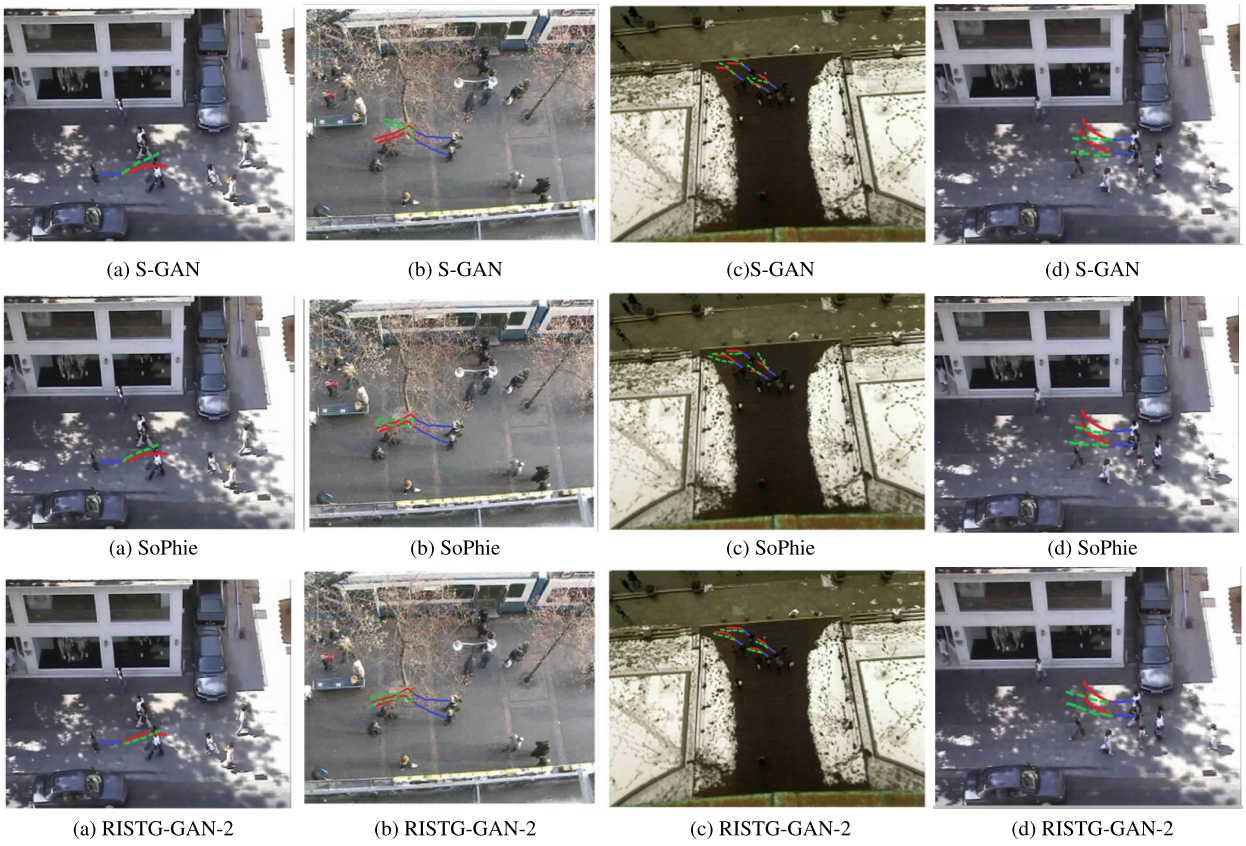


FIGURE 5. Showling the qualitative evaluation results of S-GAN, SoPhie and RISTG-GAN-2 models in typical scenes from left to right, respectively. In the picture, the observable historical and real future trajectories of pedestrians are represented by solid blue and red lines respectively, while the trajectories predicted by the model are represented by dotted green lines. a, b and c respectively represent the scenes of crossing, avoiding obstacles and following, and d is the scene of predicting failure.

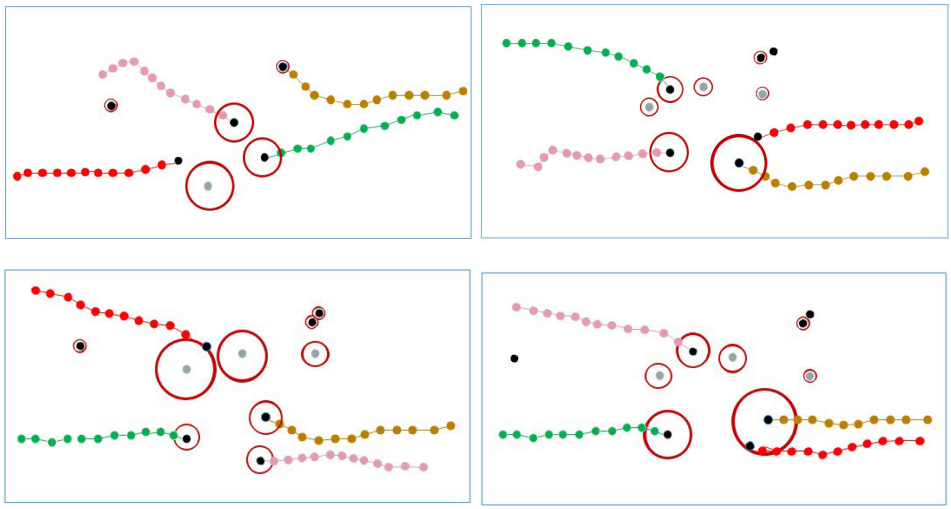


FIGURE 6. Visualization of scene attention. The red trajectory represents the predicted trajectory, and the rest are the movement trajectories of other pedestrians. The gray solid point represents the fixed obstacle position in the scene, and the black solid point represents the current moment position of the pedestrian. The red circle represents the predicted pedestrian's attention to the context of the scene, and the radius of the circle is proportional to the weight of the scene attention.

C. QUALITATIVE EVALUATION

On the basis of quantitative evaluation, we qualitatively evaluate the output prediction of S-GAN, Sophie and RISTG-GAN-2 models under four different real scenes on the ETH and UCY datasets, and the visualization results are shown in Fig. 5.

Fig. 5(a) is a crossing scene, judging from the results of qualitative evaluation, the predicted trajectories of the three models can all successfully cross the oncoming pedestrians, because the three models all model the interaction of pedestrians in the scene. However, S-GAN and Sophie only pay attention to short-term social information in the pooling process, so the prediction results are greatly different from the real trajectory. Our model, RISTG-GAN-2, uses a spatial-temporal graph to capture long-term social information, so the predicted results are closer to the real future trajectory.

Fig. 5(b) is a pedestrian interaction and avoid obstacles scene, from the evaluation results show that three kinds of models to predict the trajectory can avoid pedestrians, but S-GAN model predicts the trajectory of failed to avoid the obstacle (seat)in the scene, because the S-GAN model only consider the interaction information between the pedestrians, does not consider the scene information. The predicted trajectory of Sophie model can avoid obstacles, because the model considers the scene information, but does not consider the relative importance of the impact of obstacles on the pedestrian trajectory, so the predicted trajectory is quite different from the real trajectory. The RISTG-GAN-2 model fully considers the above-mentioned problems, so it successfully avoids obstacles, and the predicted trajectory is closer to the real future trajectory.

Fig. 5(c) is a following scene. Since the S-GAN model adopts the maximum pool mechanism and only pays attention to the most important information affecting the pedestrian trajectory, the error between the predicted trajectory and the real trajectory is the largest. The social concern component proposed by Sophie model can aggregate the information of different participants, but it is still insensitive to the unstructured characteristics of pedestrian interaction, so it also has large errors. The RISTG-GAN-2 model captures the spatial-temporal interaction information between human and environment by using spatial-temporal graph, and allocates different influence weights according to the interaction information. Therefore, the predicted trajectory of RISTG-GAN-2 is closer to the real trajectory.

Fig. 5(d) is a scene in which the prediction fails. Two pedestrians walking in a straight line change their walking direction temporarily due to a sudden vehicle passing nearby. For this situation, the predicted results of the three models are not ideal.

In Fig. 6, we visualize how much attention pedestrians pay to their surroundings. The experimental results show that people pay more attention to pedestrians and fixed obstacles in front of them than to pedestrians and fixed obstacles behind them and in the distance, which is in accordance with social common sense. Context changes behind the pedestrian or in a distant scene may affect the pedestrian's future navigation decisions.

V. CONCLUSION

In this paper, we propose an RISTG-GAN model for pedestrian trajectory prediction. The model uses spatial-temporal graph to model various interactions between human and the environment, at the same time, the time attention module is used to capture the time information of pedestrian trajectory and assign different weights. The relative importance of various interactions to pedestrian trajectory in the scene is captured by using the relative interaction scaled dot product attention module. In addition, considering the randomness of pedestrian movement in complex scene, we introduce generative adversarial network to generate the distribution of diverse trajectories in accordance with social rules. Experimental results show that our model performs better than the latest benchmark methods on multiple available datasets. Our proposed method better captures the interactions of all agents in complex scenes and improves the ability of pedestrian trajectory prediction. However, the complexity of our approach is slightly higher than that of the baseline approaches because all agents interactions are considered, but this does not affect the superiority of our approach. In the future, we will continue to optimize the model, further reduce the complexity of the model while improving the accuracy, so as to improve the navigation accuracy and real-time performance of autonomous vehicles and social robots.

REFERENCES

[1] J. Kantorovitch, J. Väre, V. Pehkonen, A. Laikari, and H. Seppälä, "An assistive household robot—Doing more than just cleaning," *J. Assistive Technol.*, vol. 8, no. 2, pp. 64–76, Jun. 2014.
[2] S. Forestier, Y. Mollard, D. Caselli, and P. Y. Oudeyer, "Autonomous exploration, active learning and human guidance with open-source poppy humanoid robot platform and explauto library," in *Proc. 30th Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, Dec. 2016, pp. 1–4.
[3] T. Bagautdinov, A. Alahi, F. Fleuret, P. Fua, and S. Savarese, "Social scene understanding: End-to-end multi-person action localization and collective activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4315–4324.
[4] W.-C. Ma, D.-A. Huang, N. Lee, and K. M. Kitani, "Forecasting interactive dynamics of pedestrians with fictitious play," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 774–782.
[5] V. Bastani, L. Marcenaro, and C. S. Regazzoni, "Online nonparametric Bayesian activity mining and analysis from surveillance video," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2089–2102, May 2016.
[6] L. Lin, Y. Lu, Y. Pan, and X. Chen, "Integrating graph partitioning and matching for trajectory analysis in video surveillance," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4844–4857, Dec. 2012.
[7] F. Jiang, Y. Wu, and A. K. Katsaggelos, "A dynamic hierarchical clustering method for trajectory-based unusual video event detection," *IEEE Trans. Image Process.*, vol. 18, no. 4, pp. 907–913, Apr. 2009.
[8] P. T. Szemes, H. Hashimoto, and P. Korondi, "Pedestrian-behavior-based mobile agent control in intelligent space," *IEEE Trans. Instrum. Meas.*, vol. 54, no. 6, pp. 2250–2257, Dec. 2005.
[9] C. Ruch, J. Gachter, J. Hakenberg, and E. Frazzoli, "The +1 method: Model-free adaptive repositioning policies for robotic multi-agent systems," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 4, pp. 3171–3184, Oct. 2020.
[10] L. Hou, L. Xin, S. E. Li, B. Cheng, and W. Wang, "Interactive trajectory prediction of surrounding road users for autonomous driving using structural-LSTM network," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4615–4625, Nov. 2020.

[11] H. Bi, Z. Fang, T. Mao, Z. Wang, and Z. Deng, "Joint prediction for kinematic trajectories in vehicle-pedestrian-mixed scenes," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, p. 10.

[12] L. Zhao, Y. Liu, A. Y. Al-Dubai, A. Y. Zomaya, G. Min, and A. Hawbani, "A novel generation-adversarial-network-based vehicle trajectory prediction method for intelligent vehicular networks," *IEEE Internet Things J.*, vol. 8, no. 3, pp. 2066–2077, Feb. 2021.

[13] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[14] B. Ivanovic and M. Pavone, "The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2375–2384.

[15] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang, "STGAT: Modeling spatial-temporal interactions for human trajectory prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6272–6281.

[16] D. Helbing and P. Molnár, "Social force model for pedestrian dynamics," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 51, p. 4282, May 1995.

[17] B. T. Morris and M. M. Trivedi, "Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2287–2301, Nov. 2011.

[18] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 961–971.

[19] N. Bisagno, B. Zhang, and N. Conci, "Group LSTM: Group trajectory prediction in crowded scenarios," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCV)*, 2018, p. 1–10.

[20] B. Douillard, D. Fox, and F. Ramos, "A spatio-temporal probabilistic model for multi-sensor multi-class object recognition," in *Robotics Research*. Cham, Switzerland: Springer, 2010, pp. 123–134.

[21] H. Koppula and A. Saxena, "Learning spatio-temporal structure from RGB-D videos for human activity detection and anticipation," in *Proc. Int. Conf. Mach. Learn. (PMLR)*, 2013, pp. 792–800.

[22] X. Zhang, P. Jiang, and F. Wang, "Overtaking vehicle detection using a spatio-temporal CRF," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2014, pp. 338–343.

[23] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-STGCNN: A social spatio-temporal graph convolutional neural network for human trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, p. 14.

[24] S. Haddad, M. Wu, H. Wei, and S. K. Lam, "Situation-aware pedestrian trajectory prediction with spatio-temporal attention model," in *Proc. 24th Comput. Vis. Winter Workshop*, 2019, pp. 1–10.

[25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[26] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Aug. 2018, pp. 2255–2264.

[27] S. Eiffert, K. Li, M. Shan, S. Worrall, S. Sukkarieh, and E. Nebot, "Probabilistic crowd GAN: Multimodal pedestrian trajectory prediction using a graph vehicle-pedestrian attention network," *IEEE Robot. Autom. Lett.*, vol. 5, no. 4, pp. 5026–5033, Oct. 2020.

[28] X. Zou, B. Sun, D. Zhao, Z. Zhu, J. Zhao, and Y. He, "Multi-modal pedestrian trajectory prediction for edge agents based on spatial-temporal graph," *IEEE Access*, vol. 8, pp. 83321–83332, 2020.

[29] J. Elfring, R. van de Molengraft, and M. Steinbuch, "Learning intentions for improved human motion prediction," *Robot. Auto. Syst.*, vol. 62, no. 4, pp. 591–602, Apr. 2014.

[30] D. Vasquez, T. Fraichard, O. Aycard, and C. Laugier, "Intentional motion on-line learning and prediction," *Mach. Vis. Appl.*, vol. 19, nos. 5–6, pp. 411–425, Oct. 2008.

[31] M. Thida, Y. L. Yong, P. Climent-Pérez, H.-L. Eng, and P. Remagnino, "A literature review on video analytics of crowded scenes," in *Intelligent Multimedia Surveillance*. 2013, pp. 17–36.

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

[33] M. T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*.

[34] X. Chen, X. Liu, Y. Qian, M. J. F. Gales, and P. C. Woodland, "CUED-RNNLM—An open-source toolkit for efficient training and evaluation of recurrent neural network language models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 1–10.

[35] H. Manh and G. Alaghband, "Scene-LSTM: A model for human trajectory prediction," 2018, *arXiv:1808.04018*.

[36] H. Xue, D. Q. Huynh, and M. Reynolds, "SS-LSTM: A hierarchical LSTM model for pedestrian trajectory prediction," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1186–1194.

[37] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[38] A. Vemula, K. Muelling, and J. Oh, "Social attention: Modeling attention in human crowds," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2017, pp. 4601–4607.

[39] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Soft + hardwired attention: An LSTM framework for human trajectory prediction and abnormal event detection," *Neural Netw.*, vol. 108, pp. 466–478, Dec. 2018.

[40] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," 2018, *arXiv:1710.10903*.

[41] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018.

[42] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," 2019, *arXiv:1909.11942*.

[43] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "SoPhie: An attentive GAN for predicting paths compliant to social and physical constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1349–1358.

[44] V. Kosaraju, A. Sadeghian, R. Martín-Martín, I. Reid, and S. Savarese, "Social-BiGAT: Multimodal trajectory forecasting using bicycle-gan and graph attention networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–10.

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[47] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 261–268.

[48] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," *Comput. Graph. Forum*, vol. 26, no. 3, pp. 655–664, 2007.

[49] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. S. Torr, and M. Chandraker, "DESIRE: Distant future prediction in dynamic scenes with interacting agents," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 336–345.

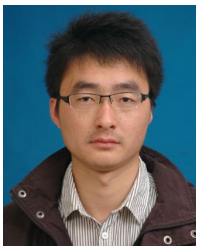
DUAN ZHAO received the B.S. and M.S. degrees from the School of Information and Electronic Engineering, China University of Mining and Technology (CUMT), Xuzhou, China, in 2006 and 2009, respectively, and the Ph.D. degree in communication engineering from the CUMT. From 2011 to 2013, he was with the School of Communication Technology, University of Duisburg–Essen, Germany, as a Visiting Ph.D. Student. He is currently a Research Assistant with the IoT Center, CUMT. His research interests include wireless sensor networks, energy harvesting in coalmine, and machine learning.

TAO LI received the B.S. degree from the College of Electronic Information Science and Technology, Xuzhou University of Technology, in 2019. He is currently pursuing the M.S. degree with the School of Information and Control Engineering, China University of Mining and Technology. His current research interests include trajectory prediction, the Internet of Things, artificial intelligence, monitoring and prediction, and edge computing.

XIANGYU ZOU received the B.S. and M.S. degrees from the School of Mechatronic Engineering, China University of Mining and Technology, Xuzhou, China, in 2008 and 2011, respectively. She is currently pursuing the Ph.D. degree with the School of Information and Control Engineering, China University of Mining and Technology. Her research interests include machine learning, edge computing, and wireless sensor networks.

YAOYI HE received the B.A. degree from the Computer Department, Xi'an Mining Institute, Xi'an, China, in 1997, and the M.A. degree from the School of Computer Technology, Nanjing University of Science and Technology, Nanjing, China, in 2008. He is currently working as a Professor with Tiandi (Changzhou) Automation Company Ltd. His research interests include the mine IoT, monitoring and control, and coal mine informatization.

LICHANG ZHAO received the B.A. degree from the College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao, China, in 2005, and the M.A. degree from the College of Information Science and Technology, University of Science and Technology of China, Hefei, China, in 2018. He is currently working as a Researcher with Tiandi (Changzhou) Automation Company Ltd. His research interests include the mine IoT, mine automation, and intellectualization.



HUI CHEN received the B.S. and M.S. degrees from the School of Mechanical and Electrical Engineering, China University of Mining and Technology (CUMT), Xuzhou, China, in 2009 and 2012, respectively. He is currently working as an Engineer with Tiandi (Changzhou) Automation Company Ltd. His main research interests include coal mine power supply and safety monitoring systems.

MINMIN ZHUO received the M.S. degree from the School of Agricultural Equipment Engineering, Jiangsu University, Zhenjiang, China, in 2018. She is currently working as an Embedded Engineer with Tiandi (Changzhou) Automation Company Ltd. Her research interests include application research of wireless technology and the development of coal mine monitoring and controlling product.