IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017. Doi Number

STI-GAN: Multimodal Pedestrian Trajectory Prediction Using Spatiotemporal Interactions and a Generative Adversarial Network

LEI HUANG[1], JIHUI ZHUANG[1], XIAOMING CHENG[1], RIMING XU [2], AND HONGJIE MA[3]

[1] Mechanical and Electrical Engineering College, Hainan University, Hainan 570228, China
[2] School of Mechanical Engineering, Beijing Institute of Technology, Beijing 100081, China
[3] Institute of Industrial Research, University of Portsmouth, Portsmouth, Hampshire PO1 2EG, United Kingdom

Corresponding author: Jihui Zhuang (e-mail: bill97@126.com).

This work was supported by the Research and Development of the China New Energy Automotive Testing Cycle Project of the Natural Science Foundation of Hainan 20155206, Hainan Province (Production Project cxy20150013) and Haikou City (Key Project 2015023).

ABSTRACT Predicting the future trajectories of multiple pedestrians in certain scenes has become a key task for ensuring that autonomous vehicles, socially interactive robots and other autonomous mobile platforms can navigate safely. The social interactions between people and the multimodal nature of pedestrian movement make pedestrian trajectory prediction a challenging task. In this paper, the problem is solved using a generative adversarial network (GAN) and a graph attention network (GAT) based on the spatiotemporal interaction information about pedestrians. Our method, STI-GAN, is based on an end-to-end GAN model that simulates the pedestrian distribution to capture the uncertainty of the predicted paths and generate more reasonable future trajectories. The complex interactions between people are modeled by a GAT, and spatiotemporal interaction information is used to improve the performance of trajectory prediction. We verify the robustness and improvement of our framework by evaluating its results on various datasets and comparing them with the results of several existing baselines. Compared with the existing pedestrian trajectory prediction methods, our method reduces the average displacement error (ADE) and final displacement error (FDE) by 21.9% and 23.8% respectively.

INDEX TERMS Pedestrian trajectory prediction, Graph attention mechanism, Generative adversarial networks, Spatiotemporal

I. INTRODUCTION

Because of its importance in video monitoring [1], planning and control of automatic driving [2], and robot navigation [3], pedestrian trajectory prediction has long been a popular focus of research in the field of computer vision. However, the prediction of pedestrian trajectories in a congested environment still presents many challenges, such as modeling the interactions between pedestrians and the surrounding environment, pedestrian trajectory uncertainty, and the capture of pedestrian intentions.

Due to the widespread application of machine learning and especially the rapid development of deep learning in recent years, researchers have mainly addressed the above challenges through related methods based on recurrent neural networks (RNNs), which serve as the background for our research. Zhu *et al.* [4] proposed an efficient method of describing interpersonal interactions through a topological star structure by observing all pedestrian trajectories and extracting a comprehensive description; however, this method ignores the impact of the surrounding environment on people. Haddad *et al.* [5] used spatiotemporal graphs to capture both the temporal and spatial correlations of pedestrian predictions and considered physical cues in a scene and the interactions between pedestrians, thereby improving the performance of trajectory prediction. In addition, Liang *et al.* [6] and Liu *et al.* [7] considered pedestrian-scene and pedestrian-object relationships simultaneously and incorporated pedestrian intentions to model future paths and predict human activities and locations. However, their work ignored the multimodal nature of the prediction of future pedestrian trajectories. As shown in Figure 1, due to the uncertainty of the future trajectories of pedestrians, compared with a multimodal trajectory prediction model, a unimodal trajectory prediction model suffers from larger errors in predicting the future trajectory distribution.

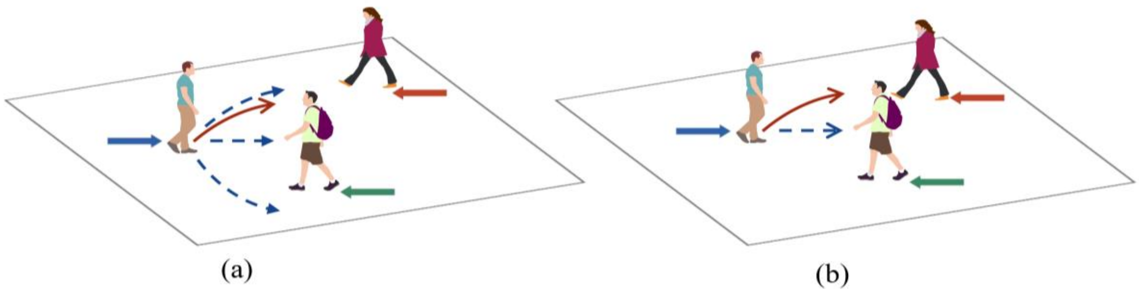


FIGURE1. **Different generative models generate different path errors when considering pedestrians walking towards each other. As shown in Fig.1(a), when a multimodal trajectory prediction model is used to predict a target pedestrian's trajectory, the pedestrian may go straight, turn left or turn right with a certain probability. This approach enables the prediction of a trajectory that is closer to the real trajectory (red trajectory) and has a smaller trajectory forecasting error. By contrast, as shown in Fig.1(b), when a unimodal trajectory prediction model is used for trajectory prediction, there may be a large error between the predicted and real trajectories of the target pedestrian. In the illustrated case, the average distance between the future trajectory of the target pedestrian (blue dashed line) and the real trajectory (red solid line) is relatively large.**

In contrast, Gupta *et al*. [8] and Amirian *et al*. [9] used a generative adversarial network (GAN) structure to model all pedestrian trajectories in a scene. These models fully consider the multimodal properties of the global scene and the trajectories, but they do not address the issue of capturing pedestrian interaction information. By comparing the global pooling and attention pooling approaches used in these two models, recent research [10] has shown that using a graph attention network (GAT) to capture pedestrian interaction information can improve the predictive performance of pedestrian interaction models.

To overcome the limitations of previous work, we propose a spatiotemporal interaction graph attention GAN (STI-GAN) model to learn the multimodal properties of the trajectories to be predicted. First, we use a graph attention network to model the social interactions of pedestrians and assign a different attention weight to each neighbor to identify neighbors of higher importance. Unlike other pooling mechanisms, the GAT allows all pedestrians in the scene to interact. Second, we implement a graph attention model based on spatiotemporal characteristics in combination with a GAN structure to generate interpretable multimodal paths in the form of end-to-end sequences and use the GAN discriminator to compare the generated paths with the real trajectories to determine how realistically the generated trajectories are. We present an experimental error analysis conducted on two publicly available real-scene pedestrian trajectory prediction datasets, and the experimental results prove the effectiveness of our proposed model.

Contributions:

1) Based on spatiotemporal information, a graph attention mechanism is extended to a GAN model to generate more accurate and interpretable multimodal path distributions.
2) Our model incorporates temporal and spatial information about social interactions to predict the future path of each pedestrian.
3) We propose an improved feature extraction method to encode the social interactions between pedestrians.

II. RELATED WORK

Our work focuses on pedestrian trajectory prediction. In the past few decades, much research has focused on traditional methods of predicting the future trajectories of pedestrians by relying on handcrafted functions [11]-[14]. Recently, however, data-driven deep learning methods have enabled great progress in this context. In this section, we discuss the existing work on RNN-related sequence prediction, graph attention network, and GAN models.

A. RECURRENT NEURAL NETWORKS (RNNS) FOR SEQUENCE PREDICTION

Pedestrian trajectory prediction is a typical sequence problem in which historical trajectory information is used to predict future trajectories. RNNs, such as long short-term memory (LSTM) networks [15] and gated recurrent unit (GRU) networks [16], are often used to process such sequence problems. In recent years, as a variant of RNNs, LSTM networks in particular have been widely used in pedestrian trajectory prediction [8], [17]-[19]. Alahi *et al.* [17] first proposed a "social pooling layer", which allows nearest-neighbor pedestrians to share hidden states, to solve interactive problems. Xue *et al.* [18] used three different LSTM networks to capture pedestrian, social, and scene size information separately and innovatively introduced factors representing the influence of the scene layout on pedestrian behavior to improve the ability to predict pedestrian trajectories. Gupta *et al.* [8] first introduced a GAN for generating multiple possible future paths for pedestrians and used a global pooling layer to accelerate the calculations. Zhang *et al.* [19] proposed an LSTM-based data-driven state refinement module, which activates the current intentions of neighbors and jointly iteratively refines the current states of all pedestrians in a crowd through a message passing mechanism.

B. GRAPH ATTENTION NETWORKS (GATS)

Recently, graph neural networks (GNNs) have been widely used in various fields, including computer vision [20], [21], recommendation systems [22], transportation networks [23], [24], and even materials chemistry [25]. The reason they are

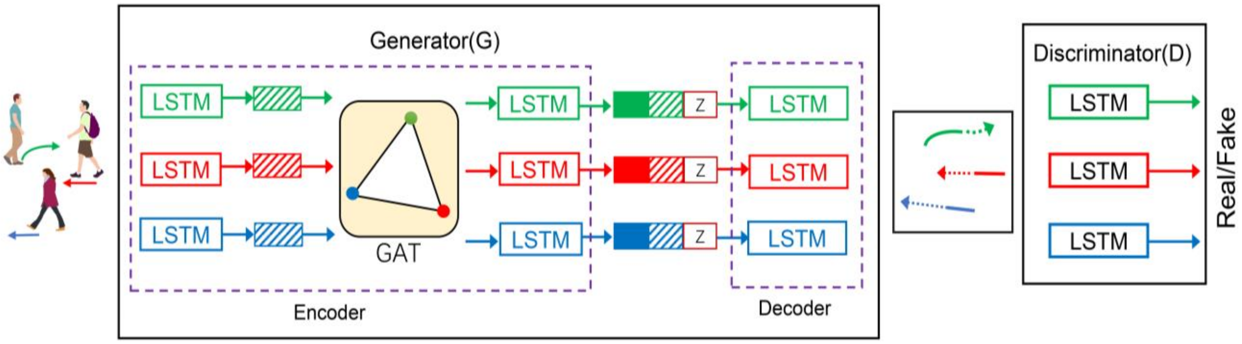


FIGURE 2. **Our proposed spatiotemporal-attention-based multimodal network architecture. The network structure is based on a GAN model and consists of three key components: a generator (G), a graph attention model (see Figure 5), and a discriminator (D). The generator uses an attention network and two LSTM modules to model the spatiotemporal correlations of interacting pedestrians and realize the fusion of spatiotemporal information.**

so widely used is that the graph structure can provide an explicit high-level representation of the environment. Networks incorporating a graph attention mechanism (graph attention networks, GATs) have also been developed based on GNNs [26]. In a GAT, different attention weights are assigned to different neighbors when aggregating feature information. Notably, the problem of pedestrian trajectory prediction has both temporal and spatial characteristics because of the changes in pedestrian movement over time and the complex interactions among different pedestrians. Accordingly, Haddad *et al*. [5] proposed an attention model based on spatiotemporal graphs that can consider the influence of surrounding pedestrians on the target pedestrians in both time and space.

In our work, we use a spatiotemporal graph model [5] and a GAT model to jointly model such complex interaction information. In each time step, we represent the interactions between pedestrians in the form of a graph, in which the pedestrians in a crowded scene correspond to the nodes of the graph and the interactions between pedestrians are described by the edges of the graph. We also assign different attention weights to different neighboring pedestrians.

C. GENERATIVE ADVERSARIAL NETWORKS (GANS)

The prediction of future pedestrian trajectories is a multimodal generation problem. Because of the capabilities of GANs in generating multimodal samples, a GAN model is suitable for solving this problem. GAN models are widely used in image translation [27], [28] and data enhancement [29]-[31] and have enabled remarkable breakthroughs in those areas. The structure of a GAN consists of a generator and a discriminator. Gupta *et al*. [8] introduced a GAN for solving the multimodal trajectory representation problem. However, the global pooling method adopted in this model uses a uniform weight for all surrounding pedestrians; thus, it can’t distinguish the different effects exerted on a target pedestrian by pedestrians at different distances and traveling at different speeds. Sadeghian *et al*. [32] improved this model by adding an attention mechanism. This improved model can assign different soft attention distribution weights to the surrounding pedestrians and the static environment, helping the model learn the interaction information of different agents and extract the most important information from the neighbors. In addition, Amirian *et al*. [9] used InfoGAN [33] to perform unsupervised learning based on data with potential categories. For the pedestrian prediction problem, our work introduces a spatiotemporal interactive encoder based on GAT that is introduced into GAN to model complex interactive behaviors in both time and space, thereby further improving the performance of trajectory prediction.

III. OUR METHOD

A. PROBLEM DEFINITION

In a scene with changing background, pedestrian position information can be obtained by an accurate target detection algorithm and be used as model input to predict the future trajectory of pedestrians. In our article, the pedestrian position is given in the dataset, and we address the prediction of the future trajectories of all pedestrians based on given pedestrian trajectories in a crowded scene. Our goal is to predict the pedestrian trajectories in future time steps $t=T_{obs+1},\dots,T_{pred}$ based on the observed trajectories $X=X_1, X_2,\dots, X_N$ of the N pedestrians in the scene in previous time steps $t=1,\dots,T_{obs.}$ The real trajectory points of pedestrian i at time t are denoted by $X_i^t=(x_i^t, y_i^t)$, and similarly, predicted future trajectory points are denoted by $\hat{Y}_i^t=(\hat{x}_i^t, \hat{y}_i^t)$.

B. OVERALL MODEL

This paper proposes a new pedestrian trajectory prediction method that can accurately predict pedestrian trajectories by comprehensively considering each pedestrian's state, movement history, and interactions with surrounding pedestrians. The network structure is shown in Figure 2. The model includes two main network components: a generator and a discriminator. The generator includes three key parts: a spatiotemporal feature coding module, a GAT module, and a decoder module.

First, the spatiotemporal feature coding module takes the historical trajectory of each pedestrian as input for feature coding and uses a combination of LSTM and GAT structures to learn the most important information about the spatiotemporal interactions between pedestrians for generating future trajectories. The learned features are then passed to the next module. The GAT module estimates the different levels of importance of the surrounding pedestrians with respect to the target pedestrian and learns the interactions between pedestrians. Subsequently, the decoder module takes the spatiotemporal interaction features along with noise as input and generates a series of reasonable future trajectories for each pedestrian. Finally, the LSTM-based discriminator compares the generated trajectories with the real trajectories and determines the probability that each generated trajectory is a real trajectory. The discriminator is mainly used to improve the predictive performance of the generator model, forcing the generator to generate more realistic samples.

C. TRAJECTORY FEATURE EXTRACTOR

The trajectory feature extraction module mainly uses an LSTM structure to extract feature representations of the observed pedestrian trajectories. We extract the nodal features of all pedestrians' past trajectories and embed the relative displacement of each pedestrian into a higher-dimensional fixed vector e_i^t through a multilayer perceptron (MLP):

$$e_i^t = MLP\left(X_i^t, W_e\right) \tag{1}$$

where W_e represents the embedding weight. Then, we use LSTM to capture the time dependence between all states of the pedestrian, for which e_i^t is used as the input to the encoder LSTM unit at time t for pedestrian i. We denote this LSTM as V-LSTM:

$$v_i^t = V\text{-}LSTM\left(v_i^{t-1}, e_i^t; W_v\right) \tag{2}$$

where v_i^t is the hidden state of the V-LSTM unit at time step t and W_v is the weight of the V-LSTM unit, which is shared among all pedestrians in the scene.

D. GAT ENCODER

GNNs are an important supplement to traditional deep learning methods because they can handle irregularly structured objects well. In this work, we extend a spatiotemporal interaction encoder based on a graph attention mechanism to a GAN. This new model can simulate the social interactions between all pedestrians in a scene from the two perspectives of spatial motion patterns and temporal correlations.

GAT and Pedestrian Construction. The "pooling" function and the "attention mechanism" mentioned in [17]-[19] cannot be used to effectively model irregularly structured objects. To model objects with irregular structures, we aggregate the information of the surrounding neighbors by adding graph attention and assigning different importance to different surrounding nodes. When calculating the spatial interaction between pedestrians in each time step, the adjacent nodes are considered mainly by introducing the GAT network, and the corresponding hidden information of each target pedestrian node is calculated and obtained. GAT introduces the self-attention mechanism to calculate the features of each neighbor node and then connects the features to obtain the influence of different neighbor nodes on the hidden state of each target pedestrian node. The GAT network is implemented by stacking multiple graph attention layers. Figure 3 shows a single graph attention layer. The input characteristic of the target node is $h = \{\vec{h}_1, \vec{h}_2, ..., \vec{h}_N\}, h \in R^F$ where N and F represent the number of nodes and characteristic dimension, respectively, and the output characteristics of nodes are $h^{'} = \{\vec{h}_1^{'}, \vec{h}_2^{'}, ..., \vec{h}_N^{'}\}, h^{'} \in R^{F^{'}}$.

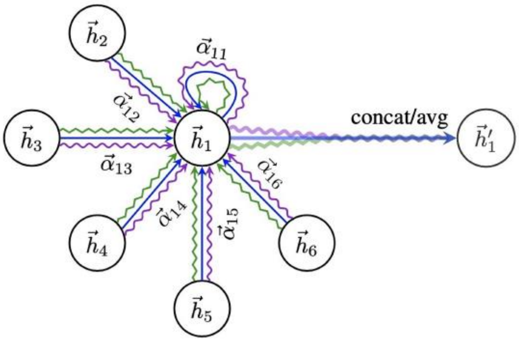


FIGURE 3. Single graph attention layer. In the graph attention layer, *K* multi attention mechanisms are applied to calculate the hidden state of nodes, and finally, their features are connected to obtain the importance of different neighbor nodes.

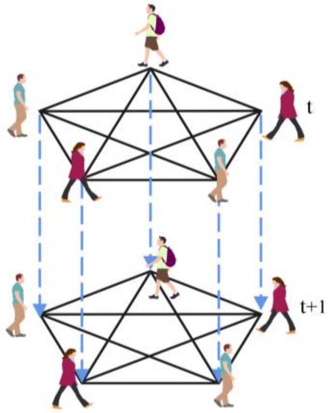


FIGURE 4. Structure diagram of pedestrian space-time interaction information in adjacent time steps. At each time step, the graphic relationship between pedestrians is represented by points and spatial edges, pedestrians are regarded as nodes, and the spatial relationship between pedestrians is represented by black solid lines. The blue directed downward dashed lines indicate temporal edges linking the same pedestrian node over adjacent time steps.

At present, our method uses GAT to model the spatial relationship between pedestrians in the same time step and uses another LSTM to capture the temporal correlation of pedestrians. Figure 4 shows the graphic structure of humans in two consecutive time steps, which mainly includes three key parts: nodes, space edges (black solid line) and time edges (blue dotted line). Among them, the nodes in the graph structure represent the pedestrians of each time step in the scene, the black solid line represents the spatial edge of the spatial social relationship between pedestrians, and the blue dotted line represents the temporal edge of the temporal correlation of the same pedestrian in the adjacent time steps.

Spatiotemporal interactive encoder based on GAT. To model the pedestrian interaction network in the crowded scene, we introduce a spatiotemporal interaction coder based on graph attention, which can model the social interaction of all pedestrians in the scene. Figure 5 describes in detail the spatiotemporal interaction input characteristics of a single node

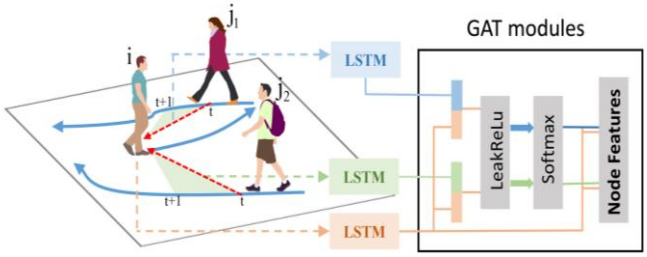


FIGURE 5. **The spatiotemporal node characteristics of pedestrians i based on the graph attention mechanism. The two red dashed lines indicate the spatial interaction between pedestrians $j_1$and j_2 to target pedestrian i at time t. The green part in front of pedestrian j_1 and pedestrian j_2 indicate the temporal influence of the surrounding pedestrians on the continuity of target pedestrian i from time t to time $t+1$. The LSTM and GAT networks are used to capture the different spatial interactions of the surrounding pedestrians on the target pedestrian, and then another LSTM is used to capture the influence of the historical trajectory of other neighbors on the target pedestrian i, that is, the temporal correlation of the motion interaction. Finally, the output g_i^t of the encoder network model is obtained as shown in formula 5.**

based on graph attention. For pedestrian i, we use the pedestrian space encoding v_i^t and v_j^t $(t=1,\dots,T_{obs})$ as the input to the softmax layer, and $\alpha_{i,j}^t$ as is used to scale the influence of the hidden state of each surrounding pedestrian $j \in N \setminus \{i\}$ on the target pedestrian. Finally, the influence of all the surrounding pedestrians is summed to form a graph and the attention layer output is g_j^t. W_g and W_u are parameters corresponding to the pedestrian. In these formulas, a denotes the shared attention mechanism, and φ is a linear embedding function:

$$u_{i,j}^t=a\left(W_u v_i^t, W_u v_j^t\right) \tag{3}$$

$$\alpha_{i,j}^t=\frac{exp\left(u_{i,j}^t\right)}{\sum_{k\in N\setminus\{i\}} exp\left(u_{i,k}^t\right)} \tag{4}$$

$$g_j^t=\sum_{j\in N\setminus\{i\}} \varphi\left(\alpha_{i,j}^t.v_j^t;W_g\right) \tag{5}$$

[10] used the hidden state for the target pedestrian as the input to the GAT. In contrast, we not only use the current historical trajectory of the target pedestrian and the spatial interactions between pedestrians in the same time step but also incorporate the historical trajectories of the other pedestrians to jointly predict their future paths.

Once the spatial interaction influence g_j^t has been obtained for each pedestrian in the crowded scene, the temporal interaction influence s_i^t on the movement of each pedestrian and the representations of the movement histories of the other pedestrians are obtained by means of another LSTM module, and we denote this LSTM module as S-LSTM. Then, we incorporate the spatial interaction influence for trajectory prediction, where || denotes a series connection and W_s denotes a parameterized shared linear transformation:

$$s_i^t=S\text{-}LSTM\left(s_i^{t-1},g_i^t;W_s\right) \tag{6}$$

$$m_i^t=(MLP_v(v_i^{T_{obs}})||MLP_s(s_i^{T_{obs}})) \tag{7}$$

E. LONG SHORT-TERM MEMORY (LSTM)-BASED GAN

As stated in the introduction, pedestrian trajectory prediction can be characterized as a multimodal problem. Accordingly, an LSTM-based GAN can be used to generate multiple reasonable trajectories. We adopt this approach to capture the uncertainty of the possible future paths.

In general, a GAN is composed of two models: a generative model and a discriminative model. The goal of the generative model is to deceive the discriminative model by generating samples that are as realistic as possible, while the goal of the discriminative model is to accurately distinguish the generated samples from the real samples. This "two-model game" ultimately enables the generative model to generate fake samples that mix the spurious with the genuine. In our model, a variety of reasonable trajectory samples are learned and predicted by a GAN.

Generator (G): The generator (G) obtains and encodes the spatiotemporal interaction information of trajectories through an LSTM-based spatiotemporal encoder and then uses an LSTM-based decoder for feature vector decoding and trajectory generation.

As shown in Figure 2, the encoder obtains the spatiotemporal interaction encoding vector m_i^t for the target pedestrian through formula 7. Following [8], the decoder takes a noise vector z sampled from a multivariate normal distribution in combination with the encoding vector representing the spatiotemporal history of a pedestrian as its input. Next, we use the LSTM method to generate the future trajectory of the pedestrian across multiple time steps $\hat{Y}_i^t\left(t=T_{obs+1},\dots,T_{pred}\right)$. We term this LSTM as G-LSTM. The corresponding LSTM model is referred to as G-LSTM. The pedestrian's future trajectory can be expressed as follows:

$$\hat{Y}_i^t=MLP_{d1}(G\text{-}LSTM((m_i^t||z),e_i^{T_{obs}};W_d),W_{d1}) \tag{8}$$

where W_d and W_{d1} are shared among all pedestrians in the scene and $e_i^{T_{obs}}$ is obtained from formula 2.

Discriminator (D): As shown in Figure 2, we use a separate encoder to learn the rules of social interaction and identify unreasonable trajectories as false. In detail, any ground-truth or generated trajectory sample may be used as the input to the discriminator, and an MLP is applied to the last hidden state of the encoder to obtain a classification score. Thus, the path is divided into a real path and a false path.

Losses: We use two different loss functions to train the network: L_{adv} and L_2. Between them, L_{adv} represents an adversarial loss, whereas L_2 is a diversity loss function applied in the trajectory generation part of the model to encourage the network to generate k different samples. The total losses are as follows:

$$L_{adv}=E[log\,D\,(X_i,Y_i)]+E\left[log\left(1\text{-}D\left(X_i,\hat{Y}_i\right)\right)\right] \tag{9}$$

$$L_2=\min_k \left\| Y_i\text{-}\hat{Y}_i^{(k)} \right\|_2 \tag{10}$$

where Y_i represents the ground-truth trajectory of pedestrian *i*. $\hat{Y}_i$ denotes the future trajectory $\hat{Y}_i$ of pedestrian *i* generated by our model, *k* is a hyperparameter, and *D* denotes the discriminator. Finally, we combine the losses to find the best discriminator $D^*$and generator G^* and choose a weight λ_1 as the final hyperparameter for combining these two loss functions:

$$G^*,D^*=\underset{G}{argmin}\underset{D}{argmax}[L_{adv}+\lambda_1 L_2] \tag{11}$$

F. IMPLEMENTATION DETAILS

In our model, an LSTM network structure is used as the RNN structure for both the generator and the discriminator. The numbers of hidden state dimensions of the generator's LSTM encoder and decoder are both 32 and that of the discriminator's LSTM encoder is 64. The input coordinates are encoded as 16-dimensional vectors and embedded into the LSTM part of the spatial encoder. During network training, only the mean square error is used for the first 250 cycles, and then, the last 250 cycles of adversarial training are conducted using both the cross-entropy loss and the mean square error. Through this training method, the generator can be encouraged to produce more reasonable results before the discriminator performs comparisons with the ground truth, thereby reducing the number of experimental iterations. In formula 8, we set $\lambda_1$to 1. During training, the Adam optimizer is used to train the generator and discriminator. The batch size is set to 64, the number of iterations is 500, and the initial learning rate is 0.01.

IV. EXPERIMENTS

In this section, the two datasets used in our experiments and the two types of prediction errors reported to evaluate the results are introduced. Then, we compare the proposed method with four other models. Quantitative and qualitative results, including results obtained by analyzing the validity of our model and visualizing the differences between trajectories, are shown.

A. DATASETS

Experiments were conducted on two public pedestrian trajectory prediction datasets: ETH [34] and UCY [35]. These two public datasets include four scenarios and five subsets: the ETH dataset includes two scenarios, namely, ETH and HOTEL, and UCY is divided into three subsets, namely, ZARA1, ZARA2, and UCY. These datasets contain 1536 pedestrians, complex social scenes, and information about the interactions between pedestrians. To make full use of the datasets when training the model, the "leave one out" method was used; i.e., the model was trained on four subsets and tested on the remaining subset. For model training, we took the first 3.2 seconds of each trajectory as the observed trajectory and predicted the trajectory over the next 3.2 seconds or 4.8 seconds. Based on the experience of the authors of S-LSTM, the data over the next 8 and 12 time steps were predicted by observing the data from the first 8 times steps, with a frame rate of 0.4 seconds.

B. BASELINES AND METRICS

Baselines. To test the effectiveness of the proposed model, we compared its performance with the performance of four other advanced models:

- **Linear:** The model is a linear regressor that estimates the linear parameters by minimizing the least square error.
- **LSTM:** The conventional LSTM model does not include a pooling mechanism, and all trajectories are considered independent of each other [15].
- **S-LSTM**: This model was proposed by Alahi *et al.* [17]. LSTM model is used to model each pedestrian. The hidden states for different pedestrians are shared between the LSTM models through a pooling mechanism. The pedestrian trajectories are predicted by modeling the interactions between different pedestrians.
- **SGAN**: This model is based on an LSTM-based codec framework that uses a GAN for training and captures the multimodal distribution of the future trajectories [8].
- **SoPhie**: This model was proposed by Sadeghian *et al.* [32]. An attention-mechanism-based GAN codec model is used to model social interactions, and a physical attention mechanism is used to achieve interpretable predictions.
- **STI-GAN**: This is the spatiotemporal multimodal GAN model proposed in this work. Following parameter settings similar to those in [8], the complete configuration of our model is denoted by STI-GAN- KV-N, where the value of *K* represents the hyperparameter used in calculating the diversity loss and the value of *N* represents the number of rounds of sampling during testing. To test the effectiveness of the model, we designed four model variants as different controls, which are represented by STI-GAN-1V-1, STI-GAN-1V-20, STI-GAN-20V-20, and SI-GAN. STI-GAN-20V-20 is our full model, and SI-GAN is the model obtained by removing the time-dependent module from STI-GAN-20V-20. STI-GAN-1V-1 means that there is no loss of diversity and the number of sampling rounds is only once; the only difference between the variant of STI-GAN-1V-20 and the complete model STI-GAN-20V-20 is the calculation of diversity loss.

Evaluation Metrics.

1) Average displacement error (**ADE**): L_2 loss between the predicted trajectory and the real trajectory on the ground averaged over all time steps $i=(1,\ldots,n)$ in the scene.

2) Final displacement error (**FDE**): The distance between the predicted final destination and the real final destination at the end of the predicted trajectory. Compared with the ADE, the FDE places more emphasis on the accuracy of destination prediction.

TABLE I. For a given trajectory over 8 times steps, the quantitative results predicted by all benchmark models over the next 8 and 12 time steps on the public datasets ETH and UCY. STI-GAN is always superior to the baseline models due to the combination of pedestrian spatiotemporal information and the graph attention mechanism in the model.

Metric	Dataset	Linear	LSTM	S-LSTM	SGAN	SoPhie	Ours (STI-GAN)			
							1V-1	1V-20	20V-20	SI-GAN
ADE	ETH	0.84/1.33	0.70/1.09	0.73/1.09	0.61/0.81	—/**0.70**	0.79/0.94	0.73/0.85	**0.54**/0.77	0.57/0.71
	HOTEL	0.35/**0.39**	0.55/0.86	0.49/0.79	0.48/0.72	—/0.76	0.39/0.82	0.34/0.78	**0.28**/0.70	0.32/0.76
	UNIV	0.56/0.82	0.36/0.61	0.41/0.67	0.36/0.60	—/0.54	0.36/0.57	**0.34/0.53**	0.35/**0.53**	0.35/0.56
	ZARA1	0.41/0.62	0.25/0.41	0.27/0.47	0.21/0.34	—/**0.30**	0.30/0.45	0.25/0.38	**0.21**/0.33	0.23/0.33
	ZARA2	0.53/0.77	0.31/0.52	0.33/0.56	0.27/0.42	—/0.38	0.27/0.40	0.22/0.34	**0.20/0.33**	0.21/0.34
AVG		0.54/0.79	0.43/0.70	0.45/0.72	0.39/0.58	—/0.54	0.42/0.64	0.38/0.58	**0.32/0.53**	0.34/0.54
FDE	ETH	1.60/2.94	1.45/2.41	1.48/2.35	1.22/1.52	—/**1.43**	1.53/1.81	1.42/1.62	**1.04**/1.53	1.10/1.44
	HOTEL	0.60/**0.72**	1.17/1.91	1.01/1.76	0.95/1.61	—/1.67	0.79/1.28	0.69/1.13	**0.54**/0.73	0.61/0.84
	UNIV	1.01/1.59	0.77/1.31	0.84/1.40	0.75/1.26	—/1.24	0.73/1.23	0.69/1.15	0.74/**1.20**	**0.73**/1.21
	ZARA1	0.74/1.21	0.53/0.88	0.56/1.00	0.42/0.69	—/**0.63**	0.61/0.96	0.50/0.83	**0.41**/0.66	0.45/0.67
	ZARA2	0.95/1.48	0.65/1.11	0.70/1.17	0.54/0.84	—/0.78	0.53/0.87	0.43/0.74	**0.41/0.66**	0.42/0.68
AVG		0.98/1.59	0.91/1.52	0.91/1.54	0.78/1.18	—/1.15	0.84/1.23	0.75/1.10	**0.63/0.96**	0.66/0.97

C. QUANTITATIVE RESULTS

1) COMPARISON WITH EXISTING WORKS.

In Table 1, our proposed model is compared with other existing models. We can see that the performance of the LSTM and S-LSTM is worse than that of SGAN and our model because GAN can effectively capture the multimodal path distribution. Besides, the proposed adversarial method based on pedestrian spatiotemporal information and a graph attention mechanism is significantly better than the previous adversarial methods [8] and [32], showing that the graph attention mechanism and the consideration of the spatiotemporal characteristics of pedestrian interactions in the model can improve its prediction performance. We also observe that SoPhie is different from other methods. It uses not only the historical paths of all agents in the scene but also scenes context information to predict the pedestrian paths. This method performs well on the ETH and ZARA2 datasets, further demonstrating the importance of considering the static scenario context for prediction. Notably, when the prediction time step is 12, in the Hotel scenario, linear performs best in both ADE and FDE. This is due to less pedestrian interaction and more linear trajectories in the Hotel scene. As shown in Table 1, the STI-GAN-20V-20 model has the smallest average error among all of the compared models. Compared with the SGAN model, its average ADEs over the next 8 and 12 time steps are reduced by 21.9% and 9.4%, respectively, and the corresponding FDEs are reduced by 23.8% and 22.9%.

2) ABLATION STUDY

Analyses were performed to evaluate the effects of the different components of the proposed model, including the diversity loss, the graph attention mechanism, and the spatiotemporal information module, as well as an evaluation of the spatial consumption. The quantitative results of different model variables are shown in the following three tables.

Evaluation of The Effect of The Diversity Loss. Due to the multimodal nature of the pedestrian movement, we generate multiple socially acceptable trajectories based on diversity loss [8]. Compared with STI-GAN-1V-1 and STI-GAN-1V-20, our final STI-GAN-20V-20 model can generate more reasonable predictions of future trajectories by means of diversity loss. The ADEs of the STI-GAN-20V-20 model for prediction over 8 and 12 future time steps are reduced by 18.8% and 19.0%, respectively, and the corresponding FDEs are reduced by 8.6% and 14.6%. The results show that the diversity loss can encourage the model to produce different predicted trajectory samples, which is helpful for improving the trajectory prediction performance of the model.

Evaluation of The Effect of The Spatiotemporal Interaction Module. To verify the effectiveness of considering spatiotemporal information, a network considering only the spatial interaction information of the crowd was also trained, that is, the SI-GAN model, which does not contain the time-dependent interaction module. As shown in Table 1, Compared with the model without the spatiotemporal interaction module SI-GAN, our full method STI-GAN-20V-20 has an ADE and FDE that are reduced by 6.3% and 4.8%, respectively, when predicting the trajectories over the next 8 times steps. This is because the spatiotemporal interaction module allows the model to consider not only the spatial interactions between pedestrians but also the influence of the continuous movement histories of the other pedestrians on the target pedestrian. The results prove that considering the spatiotemporal information of pedestrian interactions can help the model pred-

TABLE 2. Inference time (in seconds) comparison with S-LSTM and SGAN. All methods are benchmarked on the same dataset and one Tesla V100 GPU. The inference time is the average of several single inference steps.

	S-LSTM	SGAN	SI-GAN	STI-GAN
8	1.153	0.084	0.097	0.103
12	1.327	0.091	0.104	0.105
AVG	1.240	0.088	0.101	0.104
Speed-Up	14x	1x	1.15x	1.18x

TABLE 3. Comparison of CUDA memory usage. All models (S-LSTM, SGAN, SI-GAN, and STI-GAN) are benchmarked on the same dataset.

	S-LSTM	SGAN	SI-GAN	STI-GAN
Training	1059	2065	1527	1860
Validation	485	1051	600	630

ict more reasonable paths.

Evaluation of The Effect of The GAT Module. To evaluate the robustness of the graph attention mechanism, we compared two models: the SI-GAN model and the SGAN model. SI-GAN mainly uses a graph attention mechanism for modeling pedestrian interactions, while SGAN uses a pooling mechanism. From Table 1, we can see that the SI-GAN model performs slightly better than the SGAN model because the graph attention mechanism (GAT) allows the model to capture the most important pedestrian interaction information more accurately than the pooling mechanism does.

Evaluation of The Effect of The GAN Structure. To evaluate the effectiveness of the GAN discriminator, two models with different generation methods were compared: the S-LSTM model and the SI-GAN model. Between them, only the SI-GAN model relies on adversarial training to cause the output of the pedestrian trajectory prediction model, i.e., the generated distribution, to converge to the real distribution. Compared with those of the baseline S-LSTM model, the ADE and FDE of the SI-GAN model are reduced by 32.4% and 33.3%, respectively, when predicting the trajectories over the next 8 times steps, and they are reduced by 37.9% and 58.8%, respectively, when predicting the trajectories over the next 12 times steps. This is because our model uses a GAN structure to conduct adversarial training to predict reasonable future pedestrian trajectories.

Inference Speed and Spatial Consumption. The speed of pedestrian trajectory prediction is very important, for example, in practical applications such as self-driving cars and so on. The more pedestrians there are in the real scene, the more complex the graphic structure between pedestrians, and the more memory and computation required. On the public real datasets UCY and ETH, the maximum number of pedestrians per frame is 65, and the model can still accurately predict the future trajectory. Therefore, the number of pedestrians has little effect on the accuracy of trajectory prediction, but it will increase the amount of calculation.

We compared our two methods with the baseline model S-LSTM and SGAN. We refer to our complete model STI-GAN-20V-20 as STI-GAN for simplicity. As shown in Table 2, in terms of inference speed, the STI-GAN is slower than SGAN. This is because our GAT scheme is more time-consuming than SGAN's pooling module. Table 3 lists out the CUDA memory comparisons between our model and publicly available models which we could bench-mark against. The memory usage of SGAN is twice as high as that of S-LSTM during training, which indicates that adversarial training can significantly increase memory usage. We compare SI-GAN and STI-GAN indicate that considering the continuity of time interaction does not affect the speed of inference of the model, but increases the memory occupation.

D. QUALITATIVE RESULTS

In this section, we qualitatively evaluate the output predictions of SGAN, Sophie, and our complete model under four different real scenarios on the ZARA dataset. By considering spatiotemporal interaction information and a graph attention mechanism in a GAN architecture, STI-GAN can better model the relationships between pedestrians, allowing it to more accurately predict the trajectories they will follow to avoid collisions. When pedestrians walk side by side or follow each other, our model can make correct pre-dictions results. In addition, when pedestrians are walking in opposite directions, our model can better model the relationships between them to deal with such situations.

Pedestrians Walking Side by Side. On the road, it is common for pedestrians to walk side by side to the same destination while maintaining a certain distance between them. As shown in Fig. 6(a), a pair of friends walking side by side in the same direction and at the same speed. SGAN and SoPhie pay too much attention to short-term social information in the pooling process; consequently, their performance is poor. Because of the spatiotemporal interaction mechanism used in STI-GAN, however, the trajectories predicted by the STI-GAN model are roughly consistent with the real trajectories.

Person Following. On a crowded road, when a target pedestrian is following the pedestrian in front of him or her, he or she will usually keep a certain distance from the pedestrian ahead and walk in the same direction and at the same speed as that pedestrian. He or she may also deflect in a certain direction and walk forward with the pedestrian ahead. In Fig. 6(b), the trajectories of a pair of pedestrians following another pair of pedestrians. In this situation, the target pedestrian needs to pay attention to the speed and direction of the pedestrian in front and on the left and right sides at the same time. SGAN uses the maximum pooling mechanism, which only focuses on the most important features that affect pedestrian trajectories, so it generates large error prediction trajectories. STI-GAN uses its spatiotemporal interaction mechanism to aggregate and capture global pedestrian information to consider the influence of spatial relations and

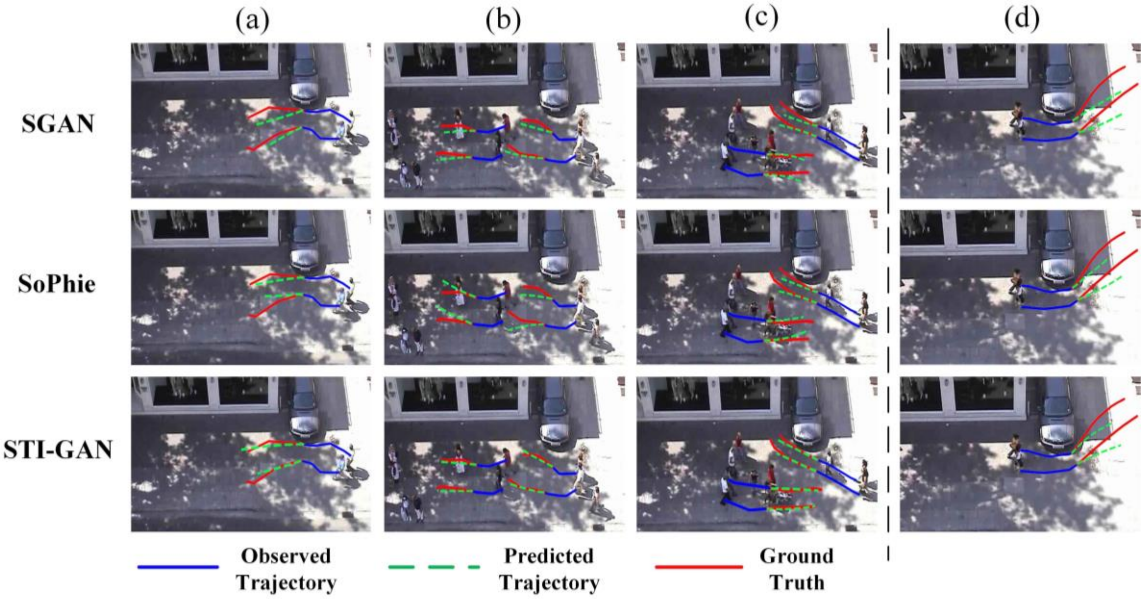


FIGURE 6. Trajectory prediction results of SGAN, SoPhie, and our proposed model in four different scenarios. Each column of images presents the trajectories predicted by the three prediction models in the same scene, i.e., the socially acceptable trajectory outputs of the different models. The blue solid lines in each image represent the historically observed trajectories, the red solid lines are the real future trajectories, and the green dotted lines are the predicted trajectories. In addition, we show some cases of prediction failure.

historical trajectories of other pedestrians on the target pedestrian. Thus, the trajectories predicted by STI-GAN are closer to the real trajectories.

Group Avoidance. When people are facing each other in a crowded scene, pedestrians usually adjust their direction and speed in time to avoid the collision between two groups of pedestrians. As shown in Fig. 6 (c), the two groups of pedestrians are facing each other, and the direction should be adjusted in time to avoid the collision between pedestrians. In this crowded environment, the key to accurate modeling is to capture the information about the interactions of the surrounding pedestrians. Among them, the predicted trajectory of SGAN is quite different from the real trajectory on the ground. SoPhie does use an attention mechanism to extract the most important trajectory information from the surrounding pedestrians, but it is still insensitive to the unstructured features of the pedestrian interactions. By virtue of the graph attention mechanism of STI-GAN, it can capture the changes in other people's intentions more successfully and learn more reliable unstructured object feature representations; and avoid collision successfully.

Failure Scenario. Another common scenario is that pedestrians suddenly change direction during the process of moving forward. Fig. 6(d) shows a pair of friends who suddenly change their direction after passing a vehicle. In this case, the prediction results of neither the proposed model nor the baseline models are ideal. SGAN shows the worst performance, while STI-GAN can better model complex pedestrian interactions by means of the graph attention mechanism and therefore still predicts trajectories that are closer to the real trajectories than SGAN does.

V. CONCLUSION

To model pedestrian motion patterns and accurately predict future pedestrian trajectories, this paper proposes a multimodal end-to-end trajectory prediction model that combines spatiotemporal interaction information based on a graph attention mechanism with the multimodal characteristics of a GAN to predict trajectories that exhibit good rationality in terms of social interactions. Our spatiotemporal graph attention model can combine spatial and temporal information to rationally assign different weights to different pedestrians in order to better capture the complex interactions between pedestrians. In addition, our GAN can produce diverse samples that conform to social rules. Our proposed model was tested on two public video datasets. The experimental results show that compared with baseline methods, the new model combining a spatiotemporal attention mechanism with a GAN can better capture the complex interactions between pedestrians to predict pedestrian trajectories in various real scenes, thereby improving the performance of pedestrian trajectory prediction.

Our work focuses on the study of social interactions between pedestrians. In the future, we can jointly model the spatiotemporal social interactions between pedestrians and other pedestrians, pedestrians and vehicles as well as vehicles and vehicles, and further improve the accuracy of trajectory prediction through joint modeling.

REFERENCES

[1] W. Xu, N. Ruiz, K. Pierce, R. Huang, J. Meyer, and J. Duthie, "Detecting pedestrian crossing events in large video data from traffic monitoring cameras," *in Proc. IEEE Int. Conf. Big Data.*, Dec. 2019, pp. 3824-3831.

[2] H. Liu, R. Xu, L. Han, and S. Xiong, "Control strategy for an electromechanical transmission vehicle based on a double markov process," *Int. J. Automot. Technol.*, (accepted).

[3] D. Verma, P. Saxena and R. Tiwari, "Robot navigation and target capturing using nature-inspired approaches in a dynamic environment," *in Proc. Conflu. Int. Conf. Cloud Comput., Data Sci. Eng.*, Jan. 2020, pp. 629-636.

[4] Y. Zhu, D. Qian, D. Ren, and H. Xia, "StarNet: Pedestrian trajectory prediction using deep neural network in star topology," in *Proc. IEEE Int Conf. Intell Rob Syst. (IROS), Nov.* 2019, pp. 8075-8080.

[5] S. Haddad, M. Wu, W. He, and S. K. Lam, "Situation-aware pedestrian trajectory prediction with spatio-temporal attention model," 2019, *arXiv :1902.05437*. [Online]. Available: https://arxiv.org/abs/ 1902.05437

[6] J. Liang, L. Jiang, J. C. Niebles, A. G. Hauptmann, and L. Fei-Fei, "Peeking into the future: Predicting future person activities and locations in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops.*, Jun. 2019, pp. 2960-2963.

[7] B. Liu, E. Adeli, Z. Cao, K. H. Lee, A. Shenoi, A. Gaidon, and J. C. Niebles, "Spatio-temporal relationship reasoning for pedestrian intent prediction," *IEEE Robot. Autom*, vol. 5, no. 2, pp. 3485-3492, Apr. 2020.

[8] 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.* (*CVPR*), Jun. 2018, pp. 2255–2264

[9] J. Amirian, J. Hayet, and J. Pettre, "Social ways: Learning multi-modal distributions of pedestrian trajectories with gans," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops.*, Jun. 2019, pp. 2964-2972.

[10] V. Kosaraju, A. Sadeghian, R. Martın-Mart´ın, I. Reid, H. Rezatofighi, and S. Savarese, "Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks," *in Proc. Adv. Neural Inf. Proces. Syst. (NIPS), Dec. 2019.*

[11] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Phys. Rev. E.*, vol. 51, no. 5, p.4282, May. 1995.

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

[13] D. Vasquez, T. Fraichard, O. Aycard, and C. Laugier, "Intentional motion online learning and prediction," *Mach. Vis. Appl.*, vol. 19, no. 5, pp. 411–425, Oct. 2008.

[14] M. Luber, J. A. Stork, G. D. Tipaldi, and K. O. Arras, "People tracking with human motion predictions from social forces," in *Proc. IEEE Int. Conf. Rob Autom. (ICRA),* May. 2010, pp. 464–469.

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

[16] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*. [Online]. Available: https://arxiv.org/abs/ 1412.3555.

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

[18] H. Xue, D. Q. Huynh, M. Reynolds, "Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.* (*WACV*), May. 2018, pp. 1186-1194.

[19] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng, "Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR),* Jun. 2019, pp. 12077–12086.

[20] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR),* Nov. 2017, pp. 3097-3106.

[21] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph r-cnn for scene graph generation," *Lect. Notes Comput. Sci.*, vol. 11205, pp. 690–706, Sep. 2018.

[22] F. Monti, M. Bronstein, and X. Bresson, "Geometric matrix completion with recurrent multi-graph neural networks," in *Proc. Adv. Neural Inf. Proces. Syst.* (*NIPS*), Dec. 2017, pp. 3697–3707.

[23] Z. Cui, K. Henrickson, R. Ke, and Y. Wang, "High-order graph convolutional recurrent neural network: a deep learning framework for network-scale traffic learning and forecasting," 2018, *arXiv :1802.07007*. [Online]. Available: https://arxiv.org/abs/ 1802.07007

[24] H. Yao, F. Wu, J. Ke, X. Tang, Y. Jia, S. Lu, P. Gong, J. Ye, and Z. Li, "Deep multi-view spatial-temporal network for taxi demand prediction," in Proc. AAAI Conf. Artif. Intell. (AAAI), Feb. 2018, pp. 2588–2595.

[25] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley, "Molecular graph convolutions: Moving beyond fingerprints," *J. Comput. -Aided Mol. Des.*, vol. 30, no. 8, pp. 595–608, Aug. 2016.

[26] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv :1609.02907*. [Online]. Available: https://arxiv.org/abs/1609.02907

[27] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *in Proc. IEEE Conf. Int. Conf. Comput.Vis.(ICCV), Oct.* 2017, pp. 2242-2251.

[28] J. Y. Zhu, R. Zhang, P. Deepak, D. Trevor, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," *in Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2017, pp. 466-477.

[29] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR),* Dec. 2018, pp. 5157-5166.

[30] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Dec*. 2018, pp. 994-1003.

[31] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR),* Dec. 2018, pp. 79-88.

[32] 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 Conf. Comput. Vis. Pattern Recognit. (CVPR),* Jun. 2019, pp. 1349–1358.

[33] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," 2016, *arXiv: 1606.03657*. [Online]. Available: https://arxiv.org/abs/1606.03657

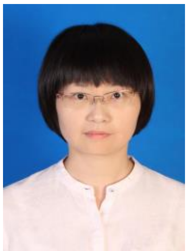
[34] S. Pellegrini, A. Ess, K. Schindler, and L. V. an Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Proc. IEEE Conf. Int. Conf. Comput.Vis.(ICCV),* Sep. 2009, pp. 261–268.

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

LEI HUANG was born in Tacheng, Xinjiang province, China in 1995. She received the B.S. degrees in Hainan University, in 2018 and she is currently pursuing the M.S. degree in Mechanical and Electrical Engineering College, Hainan University, Hainan, China. Her research interset includes pedestrian trajectory prediction, automatic driving, and image recognition.

JIHUI ZHUANG was born in Haikou, Hainan province, China in 1980. He received the B.S. degrees in Thermal energy and power engineering from Tianjin University, Tianjin, China, in 2003 and he received M.S. degree in software engineering from Tianjin University, in 2005 and the Ph.D. degree in Mechanical and power engineering from Tianjin University, in 2009.

From 2009 to 2013, he was a post-doctoral with the State Key Laboratory of internal combustion engine combustion, Tianjin University. Since 2013, he has been an associate professor with the Mechanical and Electrical Engineering College, Hainan University. He is the author of one books, more than 10 articles, and more than 10 major projects. His research interests include Key technology development of automatic driving, development of new energy vehicles, electronic control technology of engine, software and hardware development of vehicle information terminal.



XIAOMING CHENG was born in Quzhou, Zhejiang province, China in 1981.She received B.S. degrees in Thermal energy and power engineering from Tianjin University in 2003, and M.S. degrees in power machinery and engineering from Tianjin University in 2006.Her research interest includes the power unit design and matching of vehicles, and energy management of new energy vehicles.

RIMING XU was born in Shijiazhuang, Hebei province, China in 1995. He received the B.S. degrees in Hainan University, in 2019 and he is currently pursuing the M.S. degree in mechanical engineering at Beijing Institute of Technology, Beijing, China. His research interset includes the energy management of hybird vehicle and multimodal pedestrian trajectory prediction, and dynamics control of hybrid electric vehicle.

HONGJIE MA received double B.S. degrees in thermal energy and power engineering and computer science and technology from Tianjin University of Commerce, China, in 2009, and the M.S. and PhD degrees in power machinery and engineering from Tianjin University in 2015. He is a Senior Research Fellow with the School of Energy and Electronic Engineering, University of Portsmouth. He also has experience in leading the design and production of a power-train control unit and remote measurement calibration system. His research interests include data mining and artificial intelligence based diagnosis and optimization.