Multi-Modal Pedestrian Trajectory Prediction for Edge Agents Based on Spatial-Temporal Graph

XIANGYU ZOU1,2,3, BIN SUN1,2,3, DUAN ZHAO1,2,3, ZONGWEI ZHU4, JINJIN ZHAO1,2,3, AND YONGXIN HE1,2,3

1School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221008, China
2Internet of Things (Perception Mine) Research Center, China University of Mining and Technology, Xuzhou 221008, China
3National Joint Engineering Laboratory of Internet Applied Technology of Mines, Xuzhou 221008, China
4Suzhou Institute for Advanced Study, University of Science and Technology of China, Suzhou 215000, China

Corresponding author: Bin Sun (sun_bin@cumt.edu.cn)

This work was supported by the National Key Research and Development Program of China under Grant 2017YFC0804402.

ABSTRACT Edge agents, represented by socially-aware robots and autonomous vehicles, have gradually been integrated into human society. The safety navigation system in interactive scenes is of great importance to them. The key of this system is that the edge agent has the ability to predict the pedestrian trajectory in the dynamic scene, so as to avoid collision. However, predicting pedestrian trajectories in dynamic scenes is not an easy task, because it is necessary to comprehensively consider the spatial-temporal structure of human-environment interaction, visual attention, and the multi-modal behavior of human walking. In this paper, a scalable spatial-temporal graph generation adversarial network architecture (STG-GAN) is introduced, which can comprehensively consider the influence of human-environment interaction and generate a reasonable multi-modal prediction trajectory. First, we use LSTM nodes to flexibly transform the spatial-temporal graph of human-environment interactions into feed-forward differentiable feature coding, and innovatively propose the global node to integrate scene context information. Then, we capture the relative importance of global interactions on pedestrian trajectories through scaled dot product attention, and use recurrent sequence modeling and generative adversarial network architecture for common training, so as to generate reasonable pedestrian future trajectory distributions based on rich mixed features. Experiments on public data sets show that STG-GAN is superior to previous work in terms of accuracy, reasoning speed and rationality of trajectory prediction.

INDEX TERMS Trajectory prediction, spatial-temporal graph, generative adversarial network, global node, scaled dot product attention.

I. INTRODUCTION

Forecasting is a basic research topic [1]–[3], especially in pedestrian trajectory in dynamic scenes. With the rapid development of target detection [4] and edge computing [5], the deployment of many end-to-end intelligent embedded applications is accelerating. The development of a safe navigation system is a necessary step to realize the widespread integration of edge agents in an autonomous society [6]. Thanks to the breakthrough progress in detection accuracy and speed of the target detection technologies represented by SSD [7] and Yolov3 [8], by deploying these end-to-end target detection technologies, edge agents are equipped with target tracking and navigation information modeling capabilities. However, real-time target detection technology cannot support edge agents to integrate well into human society. As shown in Fig. 1, it is more beneficial for edge agents to make safe decision planning by reasoning future interactions in advance instead of reacting after the interactions occur. Therefore, pedestrian trajectory prediction in dynamic scenes is particularly important for a safe navigation system.

Pedestrian trajectory prediction in dynamic scenes is a challenging task, which requires simultaneous consideration of the spatial-temporal structure of human-environment interaction, visual attention [9] and multi-modal behavior [10] of human walking. In a dynamic scene, when making a navigation decision, the person will consider the location of the fixed obstacle and the dynamic participants in the scene to adjust their own path to avoid collision. At this moment, both fixed obstacles and dynamic participants in the scene are
treated as obstacles. The position of obstacles in the pedestrian’s field of vision at each moment can constitute a static map. With the change of time, these static maps become a dynamic map with temporal information. Therefore, the interaction between human and environment is spatiotemporal. In order to better understand interactions, humans have the ability to analyze dynamic scenes, and can quickly filter out high-value information from a large amount of information, which is the so-called attention mechanism. Inspired by the behavior and neuronal structure of the primate visual system, Itti et al. [9] proposed a visual attention model to explain this phenomenon. In addition, considering the randomness and uncertainty of human behavior, the multi-modal trajectory prediction output is more reasonable than the determined trajectory prediction.

However, most of the current pedestrian trajectory prediction methods fail to take into account the above aspects at the same time. Early trajectory prediction algorithms were based on Kalman filtering [11], Gaussian processes [12], and pattern mining [13]. Helbing and Molnar [14] pioneered the use of manual models (such as social forces) to parameterize human behavior. However, this type of model has poor adaptability and cannot solve the problem of human trajectory prediction in complex scenes. Recent methods have tried to predict pedestrian trajectories in a data-driven manner. Wong et al. [15] used the deep learning and periodic pattern mining to predict the position of the person through the semantic information of the stay points. Wu et al. [16] proposed a spatial-temporal semantic neural network algorithm that achieved relatively high accuracy in predicting human trajectories without considering interaction. However, such methods can only consider the historical trajectory of people and cannot capture the spatial-temporal interaction between pedestrians and the environment.

When considering the spatial-temporal interaction between human and environment, the spatial-temporal graph (STG) [17]–[19] has become a popular modeling tool. Because, through the spatial-temporal graph, it is possible to naturally establish the temporal and spatial connection between the target subjects, which is a necessary part of abstract modeling. Vemula et al. [20] encapsulates the social interaction in time-domain by adopting spatial-temporal graph structure, which treats social interaction as a global event occurring between each pedestrian, but does not consider scene modeling. Haddad et al. [21] proposed a long-term and short-term memory (LSTM) network based on spatial-temporal graphs, which comprehensively considered scene context information. Although such methods better model the interaction between pedestrians and the environment through spatial-temporal graphs, it does not consider the multi-modality of pedestrian trajectories, and the predicted trajectories are uniquely determined.

However, human activities are uncertain and random. Compared with the determined trajectory prediction, the multi-modal prediction trajectory distribution is more consistent with the real situation. With the great success of GAN [22] in probability calculation and behavioral reasoning, Gupta et al. [23] introduced GAN into pedestrian trajectory prediction for the first time, and proposed Social-GAN, which combined sequence prediction with generation adversarial network to generate diversified predicted trajectory without considering scene information. Zhao et al. [24] proposed Multi-Agent Tensor Fusion, which applied convolution fusion to capture the interaction of multiple agents and preserve the context of the scene, and used adversarial losses to learn random predictions without considering the relative importance of the interaction information.

The above methods either fail to take full account of the interactions in the scenario, or focus on the local impact without considering the global future impact, or only consider the equal impact of various factors without considering the relative importance of the interaction information, or fail to consider the multi-modality of the human trajectory. In order to solve the limitation of the above methods, this paper introduces a scalable spatial-temporal graph generation adversarial network architecture(STG-GAN), which can comprehensively consider the influence of human-environment interaction and generate a reasonable multi-modal prediction trajectory. First, the spatial-temporal graph of human-environment interaction is flexibly transformed into feedforward differentiable feature coding by using LSTM nodes. Different from the spatial-temporal graph in reference [20], [21], [25], we creatively propose global nodes to summarize the context changes of the scene, and integrate all node features at each moment through a forward calculation. Then, the scaled dot product attention is used to capture the relative importance of global interaction to pedestrian navigation decisions. At the same time, we use recurrent sequence modeling and generative adversarial network architecture for joint training, so as to generate a reasonable pedestrian future trajectory distribution based on the rich mixed features. The main contributions of this paper are as follows:

1) In this paper, a scalable spatial-temporal graph is proposed. The number of nodes can change dynamically according to the number of targets in the scene.
Global nodes are used to replace the paired interactions between nodes to model the global scene context. Compared with the paired interaction mode, the global node is computationally efficient.

2) In the feature encoding stage, a scaled dot product attention mechanism is introduced to capture the relative importance of global nodes to individual pedestrians, which is the real interactive impact.

3) The STG-GAN architecture proposed in this paper can generate multi-modal pedestrian trajectory prediction by using recurrent sequence modeling and generative adversarial network to train rich mixed features, which is more consistent with the characteristics of randomness of people walking in the actual scene.

The remainder of this paper is organized as follows. First, we analyze the related work of trajectory prediction. Next, we specifically explain the principle of STG-GAN pedestrian trajectory prediction model. Subsequently, we discuss the quantitative and qualitative evaluation results of the STG-GAN pedestrian trajectory prediction model on some public data sets. Finally, we summarize the work of this paper.

II. RELATED WORK
For the problem of pedestrian trajectory prediction, researchers have proposed many solutions. The existing methods can be divided into two categories: pattern-based prediction and end-to-end prediction. Fig. 2 shows the difference between the two prediction methods. Pattern-based prediction uses traditional algorithms to identify people’s motion patterns before trajectory prediction, and end-to-end prediction directly performs trajectory prediction through deep learning models. Recent research has enriched end-to-end prediction methods, including the following representative methods. For example, some scholars have introduced attention mechanisms to capture the influence of different factors on human navigation decisions in interactions, or use generative adversarial networks to model the multi-modality of human behavior.

After sorting out the previous research work, we find that the model evolution of the human trajectory prediction problem follows some rules. From hand-crafted features to data-driven, from considering a single individual to modeling social interaction and combining scene constraints, from considering only the equal impact of various factors to the introduction of attention mechanisms, from deterministic trajectory prediction to multi-modal trajectory prediction. The human trajectory prediction model is gradually improved in terms of adaptability, accuracy, and social acceptance. The related work is introduced as follows:

A. PATTERN-BASED PREDICTION
Early pattern-based methods performed trajectory prediction by mining frequent terms, constructing association rules, and sequence matching. Morris and Trivedi [26] evaluated different similarity measures and clustering methods to classify their advantages and disadvantages for trajectory learning. Jeung et al. [27] uses predefined motion functions and linear or nonlinear models to predict the future position of people. Morzy [28] uses FP-Tree Algorithm and Prefix-Span algorithm to mine frequent trajectories and motion rules, and establishes probability model of position prediction. However, this kind of research only considers the historical trajectory. These methods excavate people’s movement laws from historical trajectories, but do not consider the influence of social factors. Some literatures have pointed out that simulating social factor interaction between people can improve the accuracy and robustness of trajectory prediction models. Helbing and Molnar [14] pioneered a model of pedestrian movement with social factors, which models the attraction and repulsion between pedestrians. Pellegrini et al. [29] introduced a dynamic social behavior model that considers social interaction and the impact of different scenes on human behavior. Alahi et al. [30] proposed a social affinity map by learning the pedestrian trajectory from the relative position in the crowd to understand large-scale human mobility in specific scenes.

Although the method mentioned above has made great progress in the prediction of pedestrian trajectory, there are some limitations. These methods rely on hand-crafted behavioral model functions to capture pedestrian attributes and interaction effects in specific scenes, which results in poor model adaptability and inability to solve the problem of pedestrian trajectory prediction in complex scenes.

B. END-TO-END PREDICTION
In recent years, end-to-end recurrent neural networks (RNN) have achieved great success in sequence prediction tasks. As a variant of RNN, long short-term memory (LSTM) [31] have excellent applications in predictions with sequence characteristics such as machine translation [32], speech recognition [33], human dynamics [34]. Pedestrian trajectory prediction is also a sequence problem, and many end-to-end trajectory prediction methods have been proposed.
The current LSTM-based trajectory prediction methods can be roughly divided into two categories [35]: social-interaction methods, which simulate social interactions between moving targets, and social-scene-interaction methods, which simultaneously simulate social interactions between moving targets and scene context. Alahi et al. [36] proposed Social-LSTM, which innovatively captures pedestrian interaction through social pooling layer. Kaiping Xu et al. [37] explored the impact of different distances on pedestrians, proposed an exclusion pool layer, and expanded the Social-LSTM. However, the social-interaction method focuses on the interaction between people, but cannot understand the impact of scene context. To address this limitation, Bartoli et al. [38] proposed a Context-Aware recurrent neural network model that adds a scene-aware pooling layer to the social pooling layer. Lee et al. [39] proposed DESIRE, introducing a deep random IOCRNN encoding and decoding framework. This model uses fusion modules to jointly capture the historical trajectory of moving targets, the context of the semantic scene, and the interaction between multiple agents. Although this type of social-scene-interaction method takes into account both the social interaction and the impact of the scene, it does not consider how to measure the relative importance of various types of impacts on the trajectory of the person. For example, when people are walking on the road, they will pay more attention to the obstacles in front of them than people or vehicles in the distance. This is a way for humans to use limited attention resources to quickly screen out high-value information from a large amount of information.

In the real world, people’s navigation decisions are influenced by surrounding factors. Therefore, it is very important to measure the relative importance of various interaction effects for trajectory prediction model. Recently, some scholars have introduced attention mechanism into the field of pedestrian trajectory prediction. Vemula et al. [20] proposed a social attention mechanism that captures the relative importance of others for pedestrian navigation decisions. Haddad et al. [21] proposed the spatial-temporal attention mechanism, which is a variant of the multi-head attention mechanism, which can retain more sufficient information. Fernando et al. [40] used hard attention to weight pedestrian distances and also introduced soft attention to assess the interaction significance in the scene area. These studies show that the attention mechanism can effectively improve the accuracy of pedestrian trajectory prediction.

For the trajectory prediction problem, the above methods generate a single deterministic trajectory. However, due to the uncertainty and randomness of human activities, the multi-modal prediction trajectory is more consistent with the real situation. In the latest research [10], [23], [41], [42], researchers have considered the multi-modality of pedestrian trajectory prediction. These methods are based on generative adversarial networks, which consist of two competing networks, namely a generator and a discriminator. On the one hand, the generator tries to generate a reasonable trajectory distribution to fool the discriminator. On the other hand, the discriminator is trained to distinguish whether the generated trajectory distribution is real or fake. When the discriminator can’t distinguish the true and false tracks, it shows that the output is reasonable.

After sorting out the existing work on pedestrian trajectory prediction, we find that although some work combined spatial-temporal graph with recurrent neural network to build models, others combined convolutional neural network or recurrent neural network with generative adversarial networks to build models. However, no work has been done to combine spatial-temporal graph and generative adversarial network to solve the problem of pedestrian trajectory prediction. In this paper, first, we improve the traditional spatial-temporal graph and replace the pairwise calculation mode with a computationally efficient global node. Then, we use scaled dot product attention to capture the relative importance of global interactions to pedestrian trajectories, and innovateively propose the STG-GAN framework for joint training through sequence modeling and generative adversarial networks. This model can comprehensively consider the interaction between people and the environment and generate a reasonable multi-modal prediction trajectory. The specific principle is introduced in the third part.

III. PROBLEM REPRESENTATION AND MODEL

Before introducing the model, we first define the pedestrian trajectory prediction problem. Then, we introduce the STG-GAN encoder-decoder framework proposed in this paper. Next, we describe the working principle of feature coding using a scalable spatial-temporal graph with new global nodes. Finally, the decoder and the principle of generative adversarial networks are explained.

A. PROBLEM DEFINITION

In this paper, we focus on the prediction of dynamic pedestrian trajectories in fixed scenes (i.e. their x and y coordinates on 2D maps), taking into account the previous movement of pedestrians and the global context of the scene, including dynamic participants and fixed obstacles. Their positions at each moment can form a static map, and we assume that each scene is preprocessed to obtain the spatial coordinates of dynamic participants and fixed obstacles. The input of the STG-GAN prediction model is the observable scene context information, including the historical trajectory of all pedestrians and the location of fixed obstacles. Similar to the definition in [10], [21], [23], where the observable historical trajectory of pedestrian i is defined as: \( X_i = \{(x_{i}^t, y_{i}^t) \mid t = 1, \ldots, t_{obs}\} \), and the observable historical position of the fixed obstacle j is defined as: \( O_j = \{(x_{j}^t, y_{j}^t) \mid t = 1, \ldots, t_{obs}\} \). In addition, the real future trajectory of pedestrian i is defined as: \( Y_i = \{(x_{i}^t, y_{i}^t) \mid t = t_{obs} + 1, \ldots, t_{pred}\} \), and the predicted trajectory of pedestrian i generated by the STG-GAN model is defined as: \( \hat{Y}_i = \{(x_{i}^t, y_{i}^t) \mid t = t_{obs} + 1, \ldots, t_{pred}\} \).
B. OVERALL MODEL

In a dynamic scene, pedestrians make navigation decisions based on the movement of other dynamic participants and the limitations of fixed obstacles in the scene. In reference [21], the influence of dynamic participants and static obstacles on pedestrian trajectory is also considered, but only limited to the local interaction, which ignores the global future influence of scene context. In this paper, we model the global impact, and consider the spatial-temporal structure of human environment interaction, visual attention and the multi-modality of human walking.

To achieve this goal, a pedestrian trajectory prediction framework based on STG-GAN was proposed. The framework is mainly composed of three key modules: feature encoder module, generator / decoder module and discriminator module. In Fig. 3, the overall framework structure is shown.

In the feature encoder module, first, we use the spatial-temporal graph to establish the interaction model of dynamic participants and fixed obstacles in the scene, and extract the node feature coding from the original feature through LSTM. Then, after a forward calculation, all node features are integrated into the global node, which is the global feature coding stage. Finally, scaled dot product attention in reference [43] was used to capture the specific impact of global interaction on a single node to generate interaction feature coding.

In the generator/decoder module, we integrate random noise, node feature coding, and interaction feature coding as inputs to the generator/decoder. Based on the rich mixed features, the generator tries to generate a reasonable trajectory distribution to fool the discriminator.

In the discriminator module, the discriminator is trained to distinguish whether the generated trajectory distribution is real or fake. When the discriminator cannot clearly distinguish the generated trajectory from true, it indicates that the output is reasonable.

Based on the STG-GAN pedestrian trajectory prediction framework, we use recurrent sequence modeling and generative adversarial networks for common training to generate a reasonable pedestrian future trajectory distribution based on rich mixed features.

C. SPATIAL-TEMPORAL GRAPH ARCHITECTURE

In this paper, we use spatial-temporal graph to describe the spatiotemporal dynamic structure of human-environment interaction in the scene. Fig. 4 shows an abstract representation of the spatial-temporal graph $G = (v, \varepsilon_S, \varepsilon_T, g)$. Specifically, $G$ is a spatial-temporal graph, $v$ is a set of instance nodes, $\varepsilon_S$ is a set of spatial edges, $\varepsilon_T$ is a set of time edges, and $g$ is a newly added global node. In this paper, instance nodes include pedestrian node $P$ and obstacle node $O$, and the number of nodes varies flexibly according to the input situation. The spatial edge connects all instance nodes to the global node, and the temporal edge connects nodes with the same adjacent time step length. It should be noted that there is no need to connect the time edge between adjacent time steps of the obstacle node.

LSTM is used to encode the abstract relation described by spatial-temporal graph $G$, including spatial edge and time edge. In each time step, the corresponding LSTM shares the same parameters respectively, which enables the training network to flexibly adapt to different number of node inputs without increasing the number of parameters. This kind of
parameter sharing makes the model have the generalization ability, which can adapt to the pedestrian trajectory prediction task in the dynamic scene. The spatial-temporal graph structure in this paper adopts a variant of Structural-RNN [25]. We add global nodes to replace the original paired interaction nodes, and integrate all node features in a time step to the summary node through one-time forward calculation. In the following sections, we will introduce instance nodes, global nodes, and scaled dot product attention.

1) INSTANCE NODE
We first use Multi-Layer Perceptron (MLP) to embed the coordinate positions of each instance node separately to obtain a fixed-length vector. Taking pedestrian node \( P_i \) and fixed obstacle node \( O_j \) as examples. We embed the pedestrian node \( P_i = (x_i^t, y_i^t) \) into a vector \( e_{pi}^{i} \), which is used as the input to LSTM to get the pedestrian node coding \( h_{pi}^{i} \) (i.e. the blue square in Fig. 3). In addition, we embed the fixed obstacle node \( O_j = (x_j^t, y_j^t) \) into a vector \( e_{oj}^{i} \):

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

\[
h_{pi}^{i} = LSTM(h_{pi}^{i-1}, e_{pi}^{i}; W_{temporal}) \tag{2}
\]

\[
e_{oj}^{i} = \phi(x_j^t, y_j^t; W_o) \tag{3}
\]

where \( \phi(\cdot) \) is a non-linear embedding function, \( W_p \) and \( W_o \) are embedding weights. \( W_{temporal} \) is the weight of the spatial-temporal LSTM cell, and the weight is shared among all pedestrian nodes.

2) GLOBAL NODE
In reference [20], [21], researchers proposed to consider the interaction effect by capturing the pairwise interaction of instance nodes, which is a complex way of calculation. Assuming there are \( N \) instance nodes in the spatial-temporal graph, the above method requires \( N^2 \) calculations. In this paper, we integrate all instance node features through one forward calculation to obtain the global node, which is a key indicator describing the context of the scene. Then, all pedestrian nodes share the scene context information contained in the global node, which greatly reduces the computational overhead.

First, we use MaxPooling to integrate the coordinate position vectors of all instance nodes to obtain the global space merge vector \( g' \). Then, the vector \( g' \) is processed by a non-linear embedding function to obtain the vector \( e_{g'}^{i} \), which is used as the input of the LSTM to obtain the global feature coding \( h_{g'}^{i} \) (i.e. the orange square in Fig. 3). When considering pedestrian nodes and fixed obstacle nodes at the same time, the global feature coding \( h_{g'}^{i} \) contains all the context information of the scene; when only pedestrian nodes are considered, the global feature coding \( h_{g'}^{i} \) is degraded to include only social interaction information.

\[
g' = MaxPooling(e_{p1}^{i}, \ldots e_{pi}^{i}, e_{o1}^{i}, \ldots e_{oj}^{i}) \tag{4}
\]

\[
e_{g'}^{i} = \phi(g'; W_g) \tag{5}
\]

\[
h_{g'}^{i} = LSTM(h_{g'}^{i-1}, e_{g'}^{i}; W_{spatial}) \tag{6}
\]

where \( W_g \) is the embedding weight, \( W_{spatial} \) is the weight of the spatial-temporal LSTM cell, and the weight is shared among all global nodes.

3) SCALED DOT PRODUCT ATTENTION
According to the previous definition, the scene context features contained in global nodes are shared among all pedestrian nodes. However, the global node has different effects on different pedestrian nodes. We use a attention mechanism called scaled dot product attention to assign a global influence weight \( w(h_{g'}^{i-1}) \) to each pedestrian node, and then multiply the global influence weight \( w(h_{g'}^{i-1}) \) with the global feature coding \( h_{g'}^{i} \) to obtain the interactive feature coding \( S_{g' \rightarrow i}^{i} \) (i.e. the blue square with stripes in Fig. 3). When the global feature coding \( h_{g'}^{i} \) contains all the context information of the scene, the dot product computation can be called the scene attention mechanism. When the global feature code \( h_{g'}^{i} \) only contains social interaction information, the point product computation degenerates into a social attention mechanism similar to that in reference [20].

\[
w(h_{g' \rightarrow i}^{i}) = softmax\left(\frac{1}{\sqrt{d_e}} \text{Dot}(W_1 h_{g'}^{i}, W_2 h_{g'}^{i})\right) \tag{7}
\]

\[
S_{g' \rightarrow i}^{i} = w(h_{g' \rightarrow i}^{i}) h_{g'}^{i} \tag{8}
\]

where \( W_1 \) and \( W_2 \) are weights to linearly scale and project the hidden states into \( d_e \) dimensional vectors. \( \text{Dot}(\cdot) \) is the dot product, \( \text{softmax}(\cdot) \) is the activation function, and \( \frac{1}{\sqrt{d_e}} \) is the scale factor.

So far, the process of feature coding has been completed. We modeled the dynamic interaction process between human and environment through the spatial-temporal graph, and extracted the features that affected the pedestrian trajectory from the original features by using the LSTM module and attention mechanism. Specifically, the influencing factors are pedestrian node coding \( h_{pi}^{i} \) and interactive feature coding \( S_{g' \rightarrow i}^{i} \).
D. GENERATOR

In a classic generative adversarial network, the input to the generator is normally distributed noise. However, our goal is to obtain the reasonable future pedestrian trajectory \((\hat{x}_{t+1}, \hat{y}_{t+1})\) based on the rich feature information of human-environment interaction and the multi-modality of human walking. Therefore, we connect the pedestrian node coding \(h^t_{di}\), the interactive feature coding \(S^t_{g,si}\), and the noise vector \(z\) (i.e. the gray square in Fig. 3) in series, and use them as the input of the decoder LSTM unit to obtain the mixed feature coding \(\hat{h}^t_{di}\) (i.e. the three splited squares in Fig. 3). Then, the mixed feature coding is transformed into spatial coordinates by a multilayer perceptron. In this paper, the generator is actually the decoder.

\[
\hat{h}^t_{di} = \text{LSTM}(h^{t-1}_{di}, [h^t_{pi}, S^t_{g,si}, z]; W_d) \\
(\hat{x}^t_{t+1}, \hat{y}^t_{t+1}) = \text{MLP}(\hat{h}^t_{di}; W_{de})
\]

where \(z\) is the noise vector satisfying the standard normal distribution, MLP(\(i\)) is the multi-layer perceptron, \(W_d\) and \(W_{de}\) are the embedded weights.

E. DISCRIMINATOR

As the last part of the whole model, on the basis of observing all pedestrian historical trajectories, discriminator \(D\) will accept the real future trajectories \(Y_t\) from the test set or the generated future trajectories \(\hat{Y}_t\) from the generator, and evaluate the truth or false of the two trajectories to generate a reasonable distribution of pedestrian future trajectories.

\[
e^{t}_{disi} = \text{MLP}([X^t_i, \hat{Y}^t_i]; W_e1) \\
h^{t}_{disi} = \text{LSTM}(h^{t-1}_{disi}, e^{t}_{disi}; W_e2) \\
L_{disi} = \text{MLP}(h^{t}_{disi}; W_e3)
\]

Among them, \(\hat{Y}^t_i \sim p(Y^t_i, \hat{Y}^t_i)\) is a trajectory sample randomly selected from the real future trajectory or the predicted trajectory, and \(h^{t}_{disi}\) is the integration of \(h^{t-1}_{disi}\) at all times. \(L_{disi}\) is the result of classification. \(L_{disi} = 1\) indicates that the trajectory is real, \(L_{disi} = 0\) indicates that the trajectory is fake, \(W_{e1}\), \(W_{e2}\), \(W_{e3}\) are embedding weights.

F. LOSS

Our training goals for STG-GAN are defined as follows:

\[
V = \arg \min_{G} \max_{D} L_{GAN}(G, D) + \lambda L_{L2}(G)
\]

Among them, \(\lambda\) is the weighting coefficient, and the adversarial loss \(L_{GAN}(G, D)\) and \(L_{L2}(G)\) are defined as follows:

\[
L_{GAN}(G, D) = E_{i \in \text{STG}}[\log D(Y_i)] + E_{i \in \text{STG}}[\log (1 - D(\hat{Y}_i))]
\]

\[
L_{L2}(G) = E_{i \in \text{STG}}[\|Y_i - \hat{Y}_i\|_2]
\]

When designing the loss function objective, in addition to achieving the min-max game’s adversarial loss, we also use the \(L2\) loss to evaluate the difference between the true trajectory and the predicted trajectory.

G. IMPLEMENTATION DETAILS

In the STG-GAN pedestrian trajectory prediction framework, we did not use the convolutional neural network to extract picture features to model the scene information mentioned in the reference [24], [44]. We introduce fixed obstacle nodes (for example, trees, stone balls, street lamps, etc.), and directly abstract pedestrians and fixed obstacles as instance nodes in the spatial-temporal graph, where the coordinates of instance nodes are embedded as 16-dimensional vectors. LSTM is used as the encoder and decoder of this model. The hidden state sizes of the encoder and decoder are 16 and 32, respectively. In addition, we used the Adam optimizer [45] to train the generator and discriminator models, with an initial learning rate of 0.001. The model is trained on NVIDIA Titan XP GPU.

IV. EXPERIMENTS

In this section, we evaluate model performance on two public datasets: ETH [29] and UCY [46]. These two data sets actually record five scenes of human-environment interaction: ETH-univ, ETH-hotel, UCY-zara1, UCY-zara2 and UCY-univ. These scenes include dynamic human-environment interactions, such as pedestrians avoiding roadblocks, pedestrians crossing each other, turning at intersections, crowd following, and so on. As in [20], [23], [36], [44], we use the data within 8 seconds for model evaluation, and the recorded time step is 0.4 seconds. The first 3.2 seconds (8 time steps) are training data, and the last 4.8 seconds (12 time steps) are test data. In addition, we recorded the inference time of the model. In the model evaluation, we used two benchmark metrics, average displacement error (ADE) and final displacement error (FDE), as defined below.

ADE is defined as the average L2 distance between the true trajectory and the predicted trajectory of all pedestrians in the predicted time period.

\[
\text{ADE} = \frac{\sum_{i=1}^{N} \sum_{t=\text{obs}+1}^{\text{pred}} \|Y^t_i - \hat{Y}^t_i\|_2}{N * (t_{pred} - t_{obs} - 1)}
\]

FDE is defined as the average L2 distance between the end point of the predicted trajectory of all pedestrians and the end point of the real trajectory.

\[
\text{FDE} = \frac{\sum_{i=1}^{N} \|Y^t_i^{\text{pred}} - \hat{Y}^t_i^{\text{pred}}\|_2}{N}
\]

Among them, \(\hat{Y}^t_i\) and \(Y^t_i\) are the predicted position and real position of pedestrian \(i\) at time \(t\), and \(N\) is the number of pedestrians in the scene.

In this paper, we choose five representative models as the baselines, including LSTM [36], S-LSTM [36], SA-LSTM [20], SS-LSTM [44], and S-GAN [23]. In addition, we also performed ablation research on the STG-GAN model proposed in this paper. In the STG-GAN pedestrian trajectory prediction framework, we model human-environment interactions through flexible spatial-temporal graphs, where instance nodes are divided
into pedestrian nodes and fixed obstacle nodes. We will consider the method considering only pedestrian nodes as STG-GAN-1 and the method considering both nodes simultaneously as STG-GAN-2. In Table 1, we summarize the innovation points and modeling perspectives of these 7 models. It should be noted that only when testing the STG-GAN-2 model, the coordinates of fixed obstacles in the scene are considered, and other models cannot adapt to this information input.

### A. QUANTITATIVE EVALUATION

Table 2 lists the evaluation results of all models on the five public data sets. By observation, the prediction effect of the pure LSTM model is poor, because the model does not take into account the interaction between people and the environment at all. The evaluation effect of S-LSTM is better than the pure LSTM model because the S-LSTM model adds a social pooling layer to capture local crowd interaction to avoid collisions. The SA-LSTM model introduces Social-Attention on this basis, which refines the relative importance of crowd interaction and can better capture the pedestrian impact of different motion states in the scene. For example, in the ETH-hotel scene, there are many pedestrians staying because of waiting for a car. Social-Attention will weaken the impact of those people who stay on walking. Compared with S-LSTM, which only considers crowd interaction, the SS-LSTM model also considers the impact of local human interaction and scene characteristics. The average values of ADE and FDE on the 5 data sets were reduced by 18% and 17% respectively. This shows that the consideration of scene characteristics is helpful to improve the accuracy of trajectory prediction model.

Furthermore, compared with the LSTM-based trajectory prediction method, the GAN-based method has a smaller trajectory prediction error, which is the advantage of the multimodal prediction trajectory over the deterministic prediction trajectory. In contrast to the S-GAN model, which uses social pooling layers to capture crowd interaction information, STG-GAN-1 captures crowd interaction information through global nodes, and uses scaled dot product attention to capture the relative importance of crowd interaction to individuals. The evaluation results show that the STG-GAN-1 model further reduces the trajectory prediction error. Interestingly, although the STG-GAN-2 model considers both pedestrian nodes and fixed obstacle nodes to model scene context information. However, from the evaluation results on the three data sets UCY-zara1/zara2/univ, the STG-GAN-2 model is not much different from the STG-GAN-1 model. This may be because the fixed obstacles (vehicles parked on the side of the road) in the scene are on the edge of the road and have little effect on the navigation decision of the person. In the data sets ETH-univ and ETH-hotel, there are many fixed obstacles that need to be avoided. Considering the context information of the scene, the STG-GAN-2 model shows excellent performance in these two data sets. Compared with the STG-GAN model, the average values of the two metrics ADE and FDE of the STG-GAN-2 model on the five data sets are reduced by 17% and 13%, respectively.

In addition to evaluating the model’s prediction effect, we recorded the model’s reasoning speed in Table 3. Because inference speed is critical for the deployment of models on edge agents. The simple LSTM model has the fastest reasoning speed, but the model does not consider human-environment interaction at all, and its accuracy is poor. The SA-LSTM model uses pairwise interactive calculations to capture the relative importance of each person in crowd navigation. This calculation has the drawbacks of repeated calculations and consumes a lot of computing resources. Therefore, the model has the slowest reasoning speed. The S-LSTM and SS-LSTM models calculate social

### Table 1. Overview of innovation points and modeling perspectives of the seven baseline evaluation models.

| Model | Innovation | Modeling Perspectives |
|-------|------------|------------------------|
| LSTM  | Data driven nonlinear trajectory prediction method | Individual |
| S-LSTM| Adding pooling layer to capture local social interactions | Individual + interaction |
| SA-LSTM|Capturing the relative importance of everyone in crowd navigation | Individual + interaction + social attention |
| SS-LSTM|Considering the impact of scene information on pedestrian navigation | Individual + interaction + scene information |
| S-GAN | Considering the multi-modality of pedestrian movement | Individual + interaction + multi-modality |
| STG-GAN-1 | 1. Interactive information is modeled by spatial-temporal graphs, and instance nodes can be flexibly changed according to inputs; 2. Adding global nodes to record scene context information; 3. Using dot product attention to capture the relative importance of various impacts. | Individual + scene context + social attention + multi-modality |
| STG-GAN-2 |

### Table 2. Quantitative results of all models on 5 public data sets, separated by slashes between ADE and FDE values.

| Dataset | Baselines | STG-GAN(Ours) |
|---------|-----------|---------------|
| LSTM    | S-LSTM   | SA-LSTM | SS-LSTM | S-GAN | STG-GAN-1 | STG-GAN-2 |
| ETH-univ| 1.05/2.39| 1.09/2.35 | 0.87/1.87 | 0.92/1.92 | 0.81/1.52 | 0.73/1.38 | 0.65/1.51 |
| ETH-hotel| 0.82/1.92 | 0.79/1.76 | 0.71/1.61 | 0.67/1.52 | 0.72/1.61 | 0.69/1.42 | 0.49/1.12 |
| UCY-zara1| 0.43/0.91 | 0.47/1.00 | 0.40/0.89 | 0.38/0.75 | 0.34/0.69 | 0.31/0.64 | 0.32/0.68 |
| UCY-zara2| 0.54/1.13 | 0.56/1.17 | 0.48/0.95 | 0.41/0.82 | 0.42/0.84 | 0.38/0.81 | 0.36/0.77 |
| UCY-univ| 0.83/1.75 | 0.67/1.40 | 0.62/1.42 | 0.57/1.38 | 0.60/1.26 | 0.56/1.25 | 0.56/1.28 |
| AVG    | 0.75/1.62 | 0.72/1.54 | 0.61/1.35 | 0.59/1.27 | 0.58/1.18 | 0.54/1.10 | 0.48/1.02 |
interactions within the grid range for each pedestrian, and the calculation efficiency is relatively low. Similar to the S-GAN model that merges all global personnel interactions at one time through the social pooling layer, the STG-GAN model proposed in this paper summarizes all instance node features to the global nodes through one forward calculation. The experimental results show that the global node is more efficient than the pair computing and local grid computing to capture interactive features. Specifically, compared with the SA-LSTM model, the STG-GAN-1 inference speed is increased by 25.9 times, and the STG-GAN-2 inference speed is increased by 23.9 times.

### B. QUALITATIVE EVALUATION

On the basis of quantitative evaluation, we further study the qualitative evaluation results of typical scenes of the STG-GAN-2 model in three data sets: ETH-hotel, ETH-univ, and UCY-zara1. The visualization results are shown in Fig. 5.

In all scenarios, for pedestrians with stable motion, the multi-modal predicted trajectory distribution generated by the model is highly consistent with the pedestrian’s true future trajectory. Experimental results show that the model can better understand the following challenging scenarios, such as pedestrians avoiding obstacles, following each other, and crossing each other. Specifically, in the ETH-hotel-1/2 scenario, the model successfully predicts pedestrians to avoid trees and street lights. In the ETH-hotel-2/3 scenario, the model successfully predicts the mutual following between pedestrians. In the ETH-univ-1/2 scene, the model successfully predicts the interaction between pedestrians and the behavior of pedestrians avoiding stone balls. In summary, the model proposed in this paper has a good modeling ability for smooth human-environment interaction.

At the same time, we found that when the pedestrian motion status changes suddenly, there will be a large deviation between the predicted trajectory distribution of the

### TABLE 3. Reasoning speed comparison of seven models.

| Model         | LSTM | S-LSTM | SA-LSTM | SS-LSTM | S-GAN | STG-GAN-1 | STG-GAN-2 |
|---------------|------|--------|---------|---------|-------|-----------|-----------|
| Inference time(s) | 0.045 | 2.542  | 3.421   | 2.851   | 0.126 | 0.132     | 0.143     |
| Speed-up      | 76.1x | 1.3x   | 1x       | 1.2x     | 27.2x | 25.9x     | 23.9x     |

**FIGURE 5.** From top to bottom, the qualitative evaluation results of the STG-GAN-2 model in three data sets of ETH-hotel, ETH-univ, and UCY-zara1 are respectively. In the picture, the observable historical trajectory and the true future trajectory of the pedestrian are shown by green dashed lines and red solid lines, respectively. The yellow line distribution is the output of the diversified predicted trajectories output by the STG-GAN-2 model.
model and the true future trajectory of the pedestrian. For example, in the ETH-univ-3 scenario, the predicted output of the model is to avoid two people colliding. However, in fact, when the two met, they changed their walking directions and marched in the same direction together. In UCY-zara1-1, for two pedestrians walking in a straight line, the model cannot predict their turning intentions. In UCY-zara1-3, the abrupt walking trajectory of pedestrians near the vehicle is also unpredictable by the model. Interestingly, by observing the pedestrian trajectory prediction effect next to the car in UCY-zara1-1 and UCY-zara1-2, and the pedestrian trajectory prediction effect at the store door in UCY-zara1-2 and UCY-zara1-3. We find that as further information is input to the training model, the deviation between the predicted trajectory distribution and the true future trajectory will gradually decrease.

In addition, in Fig. 6, we visualize the attention of different pedestrians in the same scene to the context of the scene. The results show that, rather than distant pedestrians and fixed obstacles, people pay more attention to context changes in nearby scenes, which is in line with common sense. Attention to contextual changes in distant scenes may influence pedestrians’ future navigation decisions.

V. CONCLUSION

In this paper, we propose a novel pedestrian trajectory prediction model based on the STG-GAN framework, which performs better than the latest baselines on multiple public datasets. Our method completely models the various types of interactions between people and the environment through a flexible spatial-temporal graph structure. At the same time, it uses global nodes to integrate scene context information, and uses scaled dot product attention to capture the specific impact of global interactions on pedestrian trajectories. In addition, due to the uncertainty of all future movements, in order to simulate the multi-modality of pedestrian trajectories, we introduce the generative adversarial network to generate a reasonable variety of trajectories in a data-driven way. Experiments show that the STG-GAN framework proposed in this paper is accurate and efficient. It is an interesting direction to explore how edge agents can be widely integrated into human society. Next, we plan to deploy the model in combination with the edge computing platform to continue the research and we use some group intelligence algorithms [47]–[49] to optimize the LSTM model to obtain more excellent performance.

REFERENCES

[1] W. Qiao, Z. Yang, Z. Kang, and Z. Pan, “Short-term natural gas consumption prediction based on volterra adaptive filter and improved whale optimization algorithm,” Eng. Appl. Artif. Intell., vol. 87, Jan. 2020, Art. no. 103323.
[2] W. Qiao, H. Moayedi, and L. K. Foong, “Nature-inspired hybrid techniques of IWO, DA, ES, GA, and ICA, validated through a K-fold validation process predicting monthly natural gas consumption,” Energy Buildings, vol. 217, Jun. 2020, Art. no. 110023.
[3] W. Qiao, K. Huang, M. Azimi, and S. Han, “A novel hybrid prediction model for hourly gas consumption in supply side based on improved whale optimization algorithm and relevance vector machine,” IEEE Access, vol. 7, pp. 88218–88230, 2019.
[4] L. Hu and Q. Ni, “IoT-driven automated object detection algorithm for urban surveillance systems in smart cities,” IEEE Internet Things J., vol. 5, no. 2, pp. 747–754, Apr. 2018.
[5] N. Abbas, Y. Zhang, A. Taherkordi, and T. Skeie, “Mobile edge computing: A survey,” IEEE Internet Things J., vol. 5, no. 1, pp. 450–465, Feb. 2018.
[6] J. Bryson and A. Winfield, “Standardizing ethical design for artificial intelligence and autonomous systems,” Computer, vol. 50, no. 5, pp. 116–119, May 2017.
[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in Proc. Eur. Conf. Comput. Vis. Springer, 2016, pp. 21–37.
[8] J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement,” 2018, arXiv:1804.02767. [Online]. Available: http://arxiv.org/abs/1804.02767
[9] 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.

FIGURE 6. Visualization of scene attention. The red trajectory comes from the pedestrian being predicted, and the rest are the trajectories of other pedestrians. The gray rhombus indicates the fixed obstacle position in the scene, and the black solid point indicates the current position of the pedestrian. The red circle represents the predicted pedestrian’s attention to the scene context, and the radius of the circle is proportional to the scene attention weight.
BIN SUN received the B.S. degree from the College of Electronic Information Science and Technology, Xuzhou University of Technology, in 2018. 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, and edge computing.

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 CUMT. From 2011 to 2013, he was a Visiting Ph.D. Student with the School of Communication Technology, University of Duisburg-Essen, Germany. 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.

ZONGWEI ZHU received the M.S. and Ph.D. degrees in computer science from the University of Science and Technology of China (USTC), in 2011 and 2014, respectively. From 2014 to 2016, he was a Research Assistant with the IoT (Perception Mine) Research Center, China University of Mining and Technology. From 2016 to 2018, he was a Senior Engineer with Huawei Company. He is currently a Research Assistant with the Suzhou Institute, USTC. His research interests include resource scheduling, memory, power, and operating systems.

JINJIN ZHAO received the B.S. degree in information engineering from North Minzu University, in 2017. She is currently pursuing the M.S. degree with the School of Information and Control Engineering, China University of Mining and Technology. Her current research interests include smart city, the Internet of Things, artificial intelligence, and edge computing.

YONGXIN HE received the B.S. degree from the College of Electronic Information Science and Technology, Xuzhou University of Technology, in 2018. 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 the Internet of Things, wireless sensor networks, and energy harvesting technology.

* * *