**ORIGINAL ARTICLE** 



# Multi-level context-driven interaction modeling for human future trajectory prediction

Zhiquan He<sup>1,2</sup> · Hao Sun<sup>3</sup> · Wenming Cao<sup>1</sup> · Henry Z. He<sup>3</sup>

Received: 4 August 2021 / Accepted: 27 June 2022 / Published online: 18 July 2022  $\circledcirc$  The Author(s) 2022

#### Abstract

Human trajectory prediction is a challenging task with important applications such as intelligent surveillance and autonomous driving. We recognize that pedestrians in close and distant neighborhoods have different impacts on the person's decision of future movements. Local scene context and global scene layout also affect the movement decision differently. Existing methods have not adequately addressed these interactions between humans and the multi-level contexts occurring at different spatial and temporal scales. To this end, we propose a multi-level context-driven interaction modeling (MCDIM) method for human future trajectory learning and prediction. Specifically, we construct a multilayer graph attention network (GAT) to model the hierarchical human-human interactions. An extra set of long short-term memory networks is designed to capture the correlations of these human-human interactions at different temporal scales. To model the human-scene interactions, we explicitly extract and encode the global scene layout features and local context features in the neighborhood of the person at each time step and capture the spatial-temporal information of the interactions between human and the local scene contexts. The human-human and human-scene interactions are incorporated into the multi-level GAT-based network for accurate prediction of future trajectories. We have evaluated the method on benchmark datasets: the walking pedestrians dataset provided by ETH Zurich (ETH) and the crowd data provided by the University of Cyprus. The results demonstrate that our MCDIM method outperforms existing methods, being able to generate more accurate and plausible trajectories for pedestrians. The average performance gain is 2 and 3 percentage points in terms of the average displacement error and final displacement error, respectively.

Keywords Human trajectory prediction · Context modeling · Graph attention networks · Recurrent neural networks

Zhiquan He zhiquan@szu.edu.cn

> Hao Sun hshq7@mail.missouri.edu

Wenming Cao wmcao@szu.edu.cn

Henry Z. He hezhi@missouri.edu

- <sup>1</sup> Guangdong Multimedia Information Service Engineering Technology Research Center, Shenzhen University, Shenzhen 518000, China
- <sup>2</sup> Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen 518000, China
- <sup>3</sup> Video Processing and Communication Laboratory, Department of Electrical and Computer Engineering, University of Missouri, Columbia, MO 65211, USA

# 1 Introduction

Human trajectory prediction aims to predict the future moving trajectory of a person based on past observations. It has important applications in autonomous driving systems [1, 2], social robots [3, 4], human-machine interactions [5, 6], and smart environments [7, 8]. Humans can fully understand the scene situation and estimate the motion patterns of moving objects in the environment to avoid collisions. Automated vehicles also need such a capability to anticipate the pedestrians' intentions and predict their moving trajectories to reduce accidents [9]. This functionality is critical in collision avoidance systems or emergency braking systems [10]. Social robots navigating in human crowds have to move collaboratively with the surrounding people in a socially compliant way, in which accurately predicting the trajectories of surrounding humans is needed [11]. Human trajectory prediction is a challenging task since it needs to learn and predict the intentions and behaviors of humans in dynamic and complex environments. As shown in Fig. 1, a human's future trajectory is affected by many factors, including the behaviors of other persons in the scene, their social interactions, and the surrounding environment in the scene. The social norm and human intention govern the human–human interactions, such as walking in groups, observing the traffic rules, maintaining appropriate social distance, and avoiding collisions. The human–scene interactions represent the movement constraints by the physical structure, scene layout, and background motion, such as moving vehicles.

Traditional methods for human behavior modeling and prediction [12, 13] suffer from significant performance degradation with complex scenes and crowded pedestrians. The state-of-the-art methods are based on deep learning. Social-LSTM [14] and SRA-LSTM [15] used the LSTM [16] to model pedestrian movements at multiple time steps and capture the dynamic interactions among pedestrians. Its *pooling* mechanism shared the latent motion dynamics of pedestrians in the neighborhood, which were represented by the hidden states of the LSTMs. Social GAN proposed by Gupta et al. [17] and Sophie by Sadeghian et al. [18] developed generative adversarial networks (GAN) [19] to learn the distributions of multi-modal trajectories. Attention-based methods [10, 20, 21] explored graph attention networks (GAT) [22] to characterize the impact of different pedestrians in the scene on the decision of human trajectories. Huang et al. observed that both the spatial interactions at the same time step and the temporal continuity of interactions are important for predicting the future movements [20]. These methods mainly focused on humanhuman interactions. Scene-LSTM [23] considered the human-scene interactions for human trajectory prediction. Similar methods such as SS-LSTM [24], Sophie [18], Social-BiGAT [21], and Reciprocal-GAN [25] encoded the background global image features and combined them with the trajectory features to improve the trajectory prediction.

However, we recognize that the human-human interactions and the interactions between humans and multi-



**Fig. 1** A human's future trajectory in a real scenario is affected differently by different people in the neighborhood and different scene contexts. Our proposed method predicts socially and physically plausible trajectories by hierarchically modeling the influence of all involving pedestrians, the global scene layout, and the local context

level scene contexts occur at different spatial and temporal scales in real scenarios, which is not thoroughly investigated in existing works. For example, on the streets, the trajectory of a person is immediately affected by the movements of surrounding pedestrians in the local neighborhood. It is also affected by pedestrians at a further distance with whom the person will be interacting soon. For human–scene interaction, the person's trajectory is affected not only by surface conditions on the local scale (such as obstacles) but also by the overall scene structure, such as the street layout or traffic patterns on the global scale.

Motivated by this, we propose a multi-level contextdriven interaction modeling (MCDIM) method to capture these complex interactions and decisions. As illustrated in Fig. 2, our method exploits three sources of information, the human trajectory information that captures each pedestrian's past trajectories, the global scene information that extracts the scene layout from the whole scene image, and the local image patch information that captures the scene context within the local neighborhood. Specifically, we construct a multilayer graph attention network (GAT) to model the hierarchical human-human interactions, then use an extra set of LSTMs to capture the correlations of these human-human interactions at different temporal scales. To model the human-scene interactions, we explicitly extract and encode the global scene layout features and local scene context features at each time step and capture the spatial-temporal correlations of the humanscene interactions using LSTMs and GATs.

Our experimental results demonstrate that this new multi-level context-driven interaction modeling method can successfully capture the joint impact of surrounding persons and scene context on the decision of human future movement at different spatial and temporal scales. The method has significantly improved the performance of human future trajectory prediction with respect to the state of the art.

The major contribution of this work can be summarized as follows:



Fig. 2 The input information observed for our method

- 1. We address the interactions between humans and global and local scene contexts in the spatial-temporal domain and propose a multi-level context-driven interaction modeling method for human trajectory prediction.
- We construct a multilayer GAT network to model the hierarchical human-human interactions and use an extra set of LSTMs to capture the correlations of these human-human interactions at different temporal scales.
- 3. We explicitly extract and encode the image features from the global scene layout and the local scene context along the moving trajectory and model the spatial-temporal correlation of the interactions between human and local scene contexts.
- 4. We present a systematic framework to model the human-human and human-scene interactions in the spatial-temporal domain to improve the learning and prediction performance of human future trajectories.

The rest of the paper is organized as follows. Section 2 reviews the related work on human trajectory prediction and GAT. Section 3 presents our proposed multi-level context-driven interaction modeling method. Experimental results and ablation studies are presented in Sect. 4. Section 5 concludes the paper.

# 2 Related work

Our method addresses the human–human interactions and interactions between humans and multi-level scene contexts at different spatial and temporal scales using LSTMs and GATs. In this section, we review the existing works related to these two kinds of interactions. We also review the existing attention approaches for trajectory prediction, especially those based on Graph Attention Network, which is a central component of our proposed method.

# 2.1 Human-human interaction modeling for trajectory prediction

Human-human interaction modeling focuses on learning and predicting how the future trajectory of a person is affected by the surrounding pedestrians in the scene. Social Force Model proposed by Helbing et al. designed handcrafted functions to characterize pedestrian interactions using coupled Langevin equations [26]. Recently, deep neural networks, specifically recurrent neural networks, have been successfully used for human-human interaction modeling. Social-LSTM proposed a social pooling mechanism to share the hidden representations among pedestrians [14]. SRA-LSTM modeled the social interaction based on social relationship attention [15]. The key hypothesis of these works is that each pedestrian's moving direction and velocity are impacted by the surrounding pedestrians within the neighborhood. Group LSTM observed that persons tend to have coherent movement patterns [27]. Thus, it proposed to cluster the trajectories that have similar movements into groups and developed an LSTM-based model to learn group dynamics. Gupta et al. proposed an LSTM-based Generative Adversarial Network (GAN) model to learn the human-human interactions by considering all pedestrians in the scene [17]. Social-Attention [11] argued that modeling interactions between humans as a function of proximity is not necessarily true and proposed a prediction model to capture the relative importance of each individual in the scene with respect to the target person. AC-VRNN [28] used a generative Conditional Variational Recurrent Neural Network to model the human-human interactions where past observed dynamics were considered in predicting the multi-future trajectories. These methods designed various network architectures and learning algorithms to discover the complex human-human interactions but did not give enough consideration to scene context features and humanscene interactions.

# 2.2 Human-scene interaction modeling for trajectory prediction

The second category of methods considers the pedestrians' interactions with their background scenes. Zhang et al. proposed an adaptive trajectory prediction system to predict obstacle trajectories based on semantic-based dynamics and adaptively select suitable trajectory prediction methods for different types of dynamic objects [29]. SS-LSTM [24] proposed social-scene-LSTM to use a CNN (convolutional neural network) to extract the scene layout features and combine them with human-human interaction features for learning and predicting human movements. The Sophie model proposed by Sadeghian et al. [18] explored both the past trajectories of pedestrians and the semantic context of the top-view images and took two separate soft attention modules for both physical and social features to predict human trajectories. Liang et al. utilized rich visual features, such as each pedestrian's bounding box, key-point information, and scene semantic features for better prediction of multiple feasible trajectories [30]. Scene-LSTM [23, 31] designed a two-level grid structure to segment the static background scene into several cells and then trained two coupled LSTMs to encode the pedestrian's past movements and the scene structure. Matteo et al. proposed to integrate three different pooling schemes, namely social, navigation, and semantic pooling, to capture the human-human interactions, past observations from previously crossed areas, and the scene semantics, respectively [32]. This information was then fed into an LSTM-based model to predict future trajectories. Sun et al. developed the reciprocal twin networks, which include a forward prediction network to predict future trajectory from past observations and a backward prediction network that performs the trajectory prediction backward in time, to form a reciprocal constraint by utilizing the property of cycle consistency [25]. Social-BiGAT [21] considered both social and image features of the global scenes to model the human-human interactions. Most of these methods only considered the global scene information. They did not adequately address the local context around the pedestrian and the spatial-temporal correlation in the interactions between human and local scenes. We model the interactions between humans and the multi-level contexts using global and local scenes. For the global scenes, we model their hidden states using LSTMs and further apply a GAT-based attention scheme to assign adaptive weights to the scene features at different time steps.

# 2.3 Attention approaches for trajectory prediction

Attention mechanisms have been proven to be successful in various tasks. Graph neural networks (GNNs) [33] have achieved remarkable success in various machine learning tasks [34, 35]. Graph Attention Networks proposed by Velickovi et al. build upon the recent success of GNNs and incorporate self-attention into its learning process [22]. Recently, several methods applied the GATs to human trajectory prediction and have achieved state-of-the-art performance. Social-BiGAT [21] was proposed to formulate the human-human interactions as a graph and applied GATs as an attention mechanism to characterize the impact of surrounding persons on the target person's future trajectory by considering both the social and physical features of scenes. STGAT [20] shared the same problem formulation with the Social-BiGAT to model the spatial interactions between pedestrians and also designed an extra LSTM to model the temporal correlations of these interactions. Social-STGCNN [10] argued that modeling human trajectories as a graph directly from the beginning is more efficient than aggregation-based models, such as Social-BiGAT.

Unlike the existing methods reviewed above, our approach thoroughly investigates the three source factors that can affect human trajectories, i.e., the motion trajectories from surrounding people, the global scene layout, and the local scene context, and models their interactions. Furthermore, we also capture the spatial-temporal correlations of these interactions.

#### 3 Method

In this work, we recognize that the human-human interactions and the interactions between humans and multilevel scene contexts occur at different spatial and temporal scales, which are not adequately addressed in existing methods. Specifically, pedestrians in close neighborhoods and those at a far distance have different impacts on the person's decision of future movements. Local ground surface conditions and global scene layout affect the movement decision differently in the spatial-temporal domain. The following sections present a multi-level context-driven interaction modeling (MCDIM) method to address this multi-scale issue for human future trajectory learning and prediction.

#### 3.1 Problem formulation

Our problem is formulated as follows: given the observed trajectories of a group of pedestrians in a crowded scene and the contextual information of the background environment, can we predict their future trajectories? The past observed trajectories of each of the N currently visible pedestrians,  $\mathbf{X} = X_1, \dots, X_n, \dots, X_N$ , are referred to as the human trajectory information. The input trajectory for pedestrian *n* is denoted as  $X_n = (x_n^t, y_n^t)$ , where  $1 \le t \le T_o$ represents the time index (or video frame index) and  $T_o$  is the duration of observation. For a particular scene, our model extracts two parts of information for human trajectory learning and prediction. The first one is the image of the current scene  $I^t$ , which is used to capture the global layout of the current moving environment and is referred to as the global scene information. The second one is the local image block centered at the current position of the target person,  $B_n^t$ , which is used to capture the local context. Given the above observed information, our goal is to predict the future trajectories of each visible pedestrian  $n \in \{1, \cdots, N\}, \qquad \hat{Y}_n = (\hat{x}_n^t, \hat{y}_n^t)$ for time step  $t = T_o + 1, \dots, T_p$ . Here,  $T_p$  is the duration of future time for prediction. The ground truth of the future trajectories is denoted by  $Y_n$ .

#### 3.2 Method overview

An overview of our proposed multi-level context-driven interaction modeling and trajectory learning system is illustrated in Fig. 3, which mainly uses LSTMs and GATs as building blocks. LSTM has been proven to have an outstanding ability to capture the temporal correlations of the input time-series trajectories. GAT applies self-attention to a graph neural network, and it is suitable for differentially modeling the paired interactions between



Fig. 3 The pipeline of our proposed system. We take input from three sources of information: observed human trajectories, observed global scene images, and observed local scene patches. The Encoder is designed to extract spatial-temporal information from multiple

pedestrians where the importance of interaction is associated with the weights assigned to graph edges. Specifically, we first use three feature encoding modules to extract features from the past trajectories, the observed global scene images, and the local image patches around the pedestrian. Then we design three GAT-based modules to model and learn the human-human interactions, global human-scene interactions, and local human-scene interactions. We further construct extra LSTMs to capture the temporal correlations of human-human interactions and local human-scene interactions. These extra LSTMs are not applied to the global scene as the background context remains quite stationary. Finally, the hidden states for each pedestrian are then concatenated and fed into the LSTM decoder to predict the future trajectories for every pedestrian in the scene. The following section will provide a detailed explanation of these algorithm modules.

# 3.3 Spatial-temporal modeling of humanhuman interaction

The prediction of future trajectories is based on the analysis of past trajectories of all pedestrians in the scene, their interactions, and human-scene interactions. For a successful analysis of the past trajectories, we need to encode them into features. The human trajectory encoder consists of three major components: (1) A human motion extractor, (2) A GAT-based human-human interaction modeling module, and (3) An LSTM-based temporal correlation learning module, as illustrated in Fig. 4.

sources. The spatial and temporal information concatenated with noise are summarized in the intermediate State and then form the input for the Decoder to generate the predicted trajectory for each observed human



**Fig. 4** The framework of the human trajectory encoder. The upperlevel LSTMs are human motion extractors to capture the hidden motion states of each pedestrian. The middle-level GATs are used to model the human-human interactions. The lower-level LSTMs are applied to learn the temporal correlation of the interactions. Then the hidden motion states and the spatial-temporal vectors are concatenated as the output of the human trajectory encoder

#### 3.3.1 Human motion extraction and encoding

The human motion extractor is designed to capture the temporal pattern and dependency of each observed trajectory. The observed trajectory of pedestrian *n* is denoted as  $X_n = (x_n^t, y_n^t), t = 1, 2, \dots, T_o$ . First, the relative position of each pedestrian to the previous time step is calculated as follows:

$$\Delta x_n^t = x_n^t - x_n^{t-1},$$

$$\Delta y_n^t = y_n^t - y_n^{t-1}.$$
(1)

Then, we use a multilayer perceptron (MLP) network [36] to embed the relative coordinates into a fixed-size vector  $e_n^t$ :

$$e_n^t = \psi(\Delta x_n^t, \Delta y_n^t; W_e), \tag{2}$$

where  $\psi(\cdot)$  is an embedding function with ReLU nonlinearity [37] and  $W_e$  is the embedding weight.  $e_n^t$  is then fed

strategy [20, 22]. Figure 6 shows the design of one single

attention layer. The input of the graph attention layer is the hidden motion states  $M_n^t \in \{M_1^t, M_2^t, \dots, M_N^t\}$  for observed pedestrians  $1, 2, \dots, N$  at time *t*, where  $M_n^t \in \mathbb{R}^F$ , *F* is the

feature dimension of  $M_n^t$ . A new set of node features  $M_n^{\prime t} \in \mathbb{R}^{F'}$  is produced by the graph attention layer. Note that the

input and output feature dimensions, F and F', can be

different. At least one learnable linear transformation, parametrized by a weight matrix  $W_h \in \mathbb{R}^{F' \times F}$ , is applied to

obtain sufficient expressive power to transform the input features into high-level features. Then, we apply a self-

attention mechanism to the nodes to compute the attention

coefficients that indicate the importance of a node's feature

to another node. In our work, the attention mechanism is

implemented with a single layer feed-forward neural net-

work, parameterized by a weight vector  $a_h \in \mathbb{R}^{2F'}$ . The

Leaky ReLU nonlinearity, denoted by  $\Phi(\cdot)$ , is also applied.

We use the attention mechanism to compute the attention

to the LSTM network to capture the hidden motion state  $M_n^t$  for pedestrian *n* at time step *t*:

$$M_n^t = \text{LSTM}_e^h(M_n^{t-1}, e_n^t; W_e^h), \tag{3}$$

where  $LSTM_e^h$  denotes the LSTM encoder for human motion and  $W_e^h$  is the encoding weight for  $LSTM_e^h$  which is shared among all pedestrians involved in the scene and can be optimized during the training process.

#### 3.3.2 GAT-based human-human interaction modeling

When navigating in the environment, to ensure safety and to follow the rules or social norms, humans know which parts of the scene, including other pedestrians, oncoming vehicles, or special surface features, they need to pay attention to. They do not need to pay equal attention to all objects. Or, different objects in the scene will have different impacts on their decisions of future trajectories. Existing methods [14, 17, 18] use the Euclidean distance between pedestrians to weight these impact levels. We recognize that this is not necessarily true in practice. To tackle this problem, recent methods [11, 20] explore the idea of attention, where the impact levels of surrounding objects are learned and represented by an attention vector.

In order to model the human-human interactions and share information across all pedestrians in a crowded scene, we propose to explore the approach of graph attention networks by considering each pedestrian in the scene as one node in the graph. As shown in Fig. 5, we construct a complete graph at each time step, where the nodes are the pedestrians in the current scene, and the edges represent the human-human interactions. This mechanism does not introduce any restriction on pedestrian orders and allows all pedestrians to interact with each other.

The GAT network has been successfully used to analyze graph-structured data, being able to aggregate information from all neighboring nodes based on a self-attention



coefficients on each node. We then use the softmax function to normalize the attention coefficients across all nodes. After expansion, the interaction coefficients  $\alpha_{nm}^t$  of pedestrian *m* to pedestrian *n* at time step *t* are given by:  $\alpha_{nm}^t = \frac{\exp(\Phi(a_h^T[W_h M_n^t \oplus W_h M_m^t]))}{\sum_{k \in \Psi_n} \exp(\Phi(a_h^T[W_h M_n^t \oplus W_h M_k^t]))},$  (4) where  $\oplus$  denotes the concatenation operation,  $a_h^T$  represents the transposition of  $a_h$ ,  $\Psi_n$  represents the set of the neighboring nodes of node *n* on the graph. The aggregated output of the graph attention layer for pedestrian *n* at time step *t* can be computed by a linear combination of the interaction coefficients and the features corresponding to them after applying the nonlinear exponential linear unit (ELU) function  $\Theta(\cdot)$  as follows:



**Fig. 5** An illustration of the complete graph we build at each time step. Each node denotes each human  $(h_1, h_2, \dots, h_n)$ , and the edges represent the human–human interaction

**Fig. 6** An illustration of the computing mechanism of a single graph attention layer. It aggregates information from each neighboring node and follows a self-attention strategy.  $M_1^t, M_2^t, \dots, M_N^t$  indicates the hidden motion states for pedestrians  $1, 2, \dots, N$ , and  $\alpha_{11}, \alpha_{12}, \dots, \alpha_{1N}$  denotes the importance of the corresponding pedestrian with respect to pedestrian 1

$$M_n'^t = \Theta\left(\sum_{m \in \Psi_n} \alpha_{nm}^t W_h M_m^t\right).$$
<sup>(5)</sup>

# 3.3.3 Correlation analysis for human-human interaction modeling at multiple temporal scales

By far, we have modeled human-human interactions by sharing hidden states among pedestrians at the same time step. However, these operations only capture the spatial information of human-human interactions. In order to capture the human-human interactions at different temporal scales, we design an extra set of LSTMs for learning the temporal correlations between those interactions. Here, we use the LSTMs because they can learn complex non-linear dependencies and correlation within time-series data [20]. The input of the temporal LSTM (LSTM<sup>h</sup><sub>t</sub>) is the output of the graph attention layers  $M'_n^t$  from (5). The operation of LSTM<sup>h</sup><sub>t</sub> is described by

$$T_n^t = \text{LSTM}_t^h(T_n^{t-1}, M_n^{\prime t}; W_t^h), \tag{6}$$

where  $T_n^t$  is the hidden temporal correlation state of human-human interactions and  $W_t^h$  is the weight for LSTM<sub>t</sub><sup>h</sup> and shared among all sequences. In the last step of the human trajectory encoder, we concatenate the hidden motion states  $M_n^t$  obtained from (3) and the temporal correlations of human-human interactions  $T_n^t$  obtained from (6) to form the spatial-temporal information learned from the observed human trajectories:

$$H_e = M_n^t \oplus T_n^t. \tag{7}$$

# 3.4 Multi-level human-scene interaction modeling

As discussed in Sect. 1, a person's decision and choice of future trajectory are impacted by the navigation scene context and human–scene interactions at different spatial– temporal scales.

#### 3.4.1 Modeling the impact of global scene context

SS-LSTM [24] and Sophie [18] have recognized that the chosen trajectory of a pedestrian is affected by the scene layout, such as stationary obstacles, moving cars, entries, and exits. In our work, we propose to extract global information of the scene and use a GAT network to model the global human–scene interactions. Fig. 7 illustrates the design of our global scene layout encoder network which is used to extract and encode the global scene layout



**Fig. 7** The framework of global scene layout encoder. The inputs are the observed scene images at each time step. A CNN is used to extract the scene layout features, which are then fed to LSTMs to compute the hidden state vectors. A GAT is applied to model the global human–scene interactions

information. First, we use a CNN to extract the global scene layout features  $l^t$  from scene image  $I^t$  at each time step *t*:

$$t^{t} = \text{CNN}(I^{t}). \tag{8}$$

In our experiments, we choose the VGGNet-19 network [38] that is pre-trained on the ImageNet [38]. The extracted scene layout feature  $l^t$  is then fed to an LSTM encoder (LSTM<sub>e</sub><sup>l</sup>) to compute the hidden state vector  $L^t$  for each observed pedestrian at time step *t*:

$$L^{t} = \mathrm{LSTM}_{e}^{l} \left( L^{t-1}, l^{t}; W_{e}^{l} \right), \tag{9}$$

where  $W_e^l$  is the associated encoding weight [24].

Since the camera is stationary in our task, the background scene remains almost the same in most cases. However, the pedestrians in the scene are moving. The configuration of the observed scene images also changes with time due to the movement of these pedestrians or background objects, such as parked or moved vehicles. Therefore, the difference between scene features at different time steps is mainly caused by the background motion. In order to model these human-scene interactions, we propose to use a GAT network to assign different and adaptive attention weights to the hidden states of the scene features at different time steps. Specifically, we first construct a complete graph by considering each hidden state of the scene features as graph nodes and their interactions as edges, as shown in Fig. 6. The inputs to graph attention layers are the hidden states of the scene features at each time step of  $L^1, L^2, \dots, L^{T_o}$ . The interaction coefficients  $\alpha_{kt}$ between the scene hidden states  $L^t$  and  $L^k$  are defined as follows:

$$\alpha_{kt} = \frac{\exp(\Phi(a_g^T[W_gL^t \oplus W_gL^k]))}{\sum_{k \in \Psi_{T_o}} \exp(\Phi(a_g^T[W_gL^t \oplus W_gL^k]))}.$$
(10)

Similar to (4),  $a_g$  is the weight vector of a single layer feedforward neural network that is normalized by a softmax function with Leaky ReLU denoted by  $\Phi(\cdot)$ .  $a_g^T$  is the transpose of  $a_g$ .  $\Psi_{T_o}$  represents the set of the scene feature hidden states from time step 1 to  $T_o$ .  $W_g$  is the weight matrix of a shared linear transformation that is applied to each node. The aggregated output of the graph attention layer for scene feature hidden state at time step t can be computed by

$$L'^{t} = \Theta\left(\sum_{k \in \Psi_{T_{o}}} \alpha_{kt} W_{g} L^{t}\right), \tag{11}$$

where  $\Theta(\cdot)$  is a nonlinear function, *e.g.*, exponential linear unit (ELU) function [20]. Finally, the global scene layout hidden state and global human–scene interaction are concatenated together to form the output of the global scene layout encoder for each pedestrian:

$$G_e = L^t \oplus {L'}^t, \tag{12}$$

where  $\oplus$  indicates the concatenation operation.

#### 3.4.2 Modeling the impact of local context

Another important factor in determining the future trajectory is the local context. The ground surface condition in the local neighborhood, such as obstacles and potential collision with incoming persons or vehicles, has a direct impact on the pedestrian's decision of future trajectory. Note that this local context is evolving as the person is walking. To capture this local context and analyze its impact on the person's decision on the future trajectory, we propose to extract a sequence of local image patches around the moving person, encode them into features, and incorporate them into the GAT-based trajectory learning and prediction framework, as illustrated in Fig. 8. The local patch for pedestrian *n* at time step *t* is denoted as  $B_n^t$ , which is segmented from the scene image with a fixed size. For example, in our experiments, we set the patch size to  $128 \times 128$ . To encode the local context image patches, we use a pre-trained CNN to extract the context features  $c_n^t$  of the local patch  $B_n^t$ :



**Fig. 8** The framework of local patch context encoder. The inputs are the observed certain-sized local image patches centered on the position of each pedestrian at each time step. A CNN is used to extract the local context features, which are then fed to LSTMs to capture the hidden state of the local context. A GAT is applied to model the local human–scene interactions. An extra set of LSTMs is designed to learn the temporal correlations of these interactions. The tensors from the upper-level LSTMs and the lower-level LSTMs are concatenated to form the output of the local patch context encoder

$$c_n^t = \text{CNN}(B_n^t). \tag{13}$$

These features are then fed into the LSTMs to capture the hidden state of the local context:

$$C_n^t = \text{LSTM}_e^c(C_n^{t-1}, c_n^t; W_e^c), \qquad (14)$$

where  $W_{\rho}^{c}$  is the corresponding weight for LSTM<sub>e</sub><sup>c</sup>.

To incorporate the local scene context into the GAT analysis framework, we construct a complete graph with the hidden state of local context for each pedestrian at the same time step as graph nodes and their interactions as the edges. The interaction weight  $\beta_{nm}^t$  between the local patch  $B_m^t$  and  $B_n^t$  at time step t is computed by

$$\beta_{nm}^{t} = \frac{\exp(\Phi(a_l^T[W_l C_n^t \oplus W_l C_m^t]))}{\sum_{k \in \Psi_n} \exp(\Phi a_l^T[W_l C_n^t \oplus W_l C_k^t]))},\tag{15}$$

where  $W_l$  is the weight matrix of a shared linear transformation that is applied to each node,  $a_l$  is the weight vector of a single layer feed-forward neural network normalized by a softmax function with Leaky ReLU  $\Phi(\cdot)$ ,  $a_l^T$  is the transpose of  $a_l$ ,  $\oplus$  represents the concatenation operation,  $\Psi_n$  represents the set of the neighboring nodes of node *n* on the graph. With the normalized attention coefficients  $\beta_{nnn}^t$ , the aggregated output of the graph attention layer for pedestrian *n* at time step *t* is computed as follows:

$$C_n^{\prime t} = \Theta(\sum_{m \in \Psi_n} \beta_{nm}^t W_l C_m^t), \tag{16}$$

where  $\Theta(\cdot)$  is a nonlinear function, specifically, the exponential linear unit (ELU) function [20].

#### 3.4.3 Temporal correlation learning local human-scene interactions

As in the human-human interaction modeling, the humanscene interactions also occur at different temporal scales with different correlation patterns. To address this issue, we also design an extra set of temporal LSTMs to learn the temporal correlation between the local human-scene interactions modeled at the same time step:

$$P_{n}^{t} = \text{LSTM}_{t}^{c} \left( P_{n}^{t-1}, C_{n}^{\prime t}; W_{t}^{l} \right), \tag{17}$$

where  $P_n^t$  is the hidden temporal correlation state of local human-scene interactions  $C'_n^t$ , and  $W_t^l$  is the associated weight. Both the local scene context of the target pedestrian and the local human-scene interactions are useful to help us better learn human behavior. By taking advantage of the spatial-temporal information, we combine these two parts together, which is given by

$$L_e = C_n^t \oplus P_n^t, \tag{18}$$

where  $\oplus$  is the concatenation operation.

#### 3.5 Future trajectory prediction

As illustrated in Fig. 3, after the human-human and human-scene interactions at different spatial-temporal scales have been encoded, the intermediate state tensor  $S_n^t$  for each pedestrian is formed by concatenating the hidden states of the pedestrian motion  $H_e$  from (7), the global scene layout  $G_e$  from (12), the local context  $L_e$  from (18), and a random noise Z with Gaussian distribution  $\mathcal{N}(0, 1)$ :

$$S_n^t = H_e \oplus G_e \oplus L_e \oplus Z. \tag{19}$$

The purpose of Gaussian noise Z will be explained in the following. Then the intermediate state tensor is fed into the LSTM decoder as the initial hidden state to predict the future relative coordinates Y for each pedestrian by

$$\hat{S}_n^t = \text{LSTM}_d(\hat{S}_n^{t-1}, S_n^t; W_d), 
Y_t^n = (\hat{x}_n^t, \hat{y}_n^t) = W_o \hat{S}_n^t + b_o,$$
(20)

where  $W_d$  is the weight matrix of the LSTM decoder,  $W_o$  and  $b_o$  are the corresponding weight and bias term of the linear output layer. The relative coordinates then can be easily converted to the real coordinates according to (1).

Similar to the previous methods [17, 20], given the observation of past human trajectories, our model aims to learn human motion patterns and generate multiple both physically and socially feasible trajectories. Methods in [14, 24, 31] proposed to generate the future trajectory by sampling from a Gaussian distribution, where the parameters for the distribution are trainable by minimizing the negative log-likelihood loss of the ground truth. However, as discussed in [17], the sampling process is not differentiable, which does not allow an end-to-end learning process.

Thus, in our work, we follow the strategy in previous works [17, 20] to model the multi-modal property of human movements by introducing a variety loss to encourage diversity of generated trajectories from the network. Specifically, for each pedestrian, k possible trajectories are generated by introducing a random noise  $Z \sim \mathcal{N}(0, 1)$  before the decoder stage. Then, the predicted trajectory, which has the smallest distance to the ground truth, is chosen as the final output. The variety loss is given by

$$L_{\text{variety}} = \min_{k} ||Y_n - \hat{Y}_n^t||_2, \qquad (21)$$

where  $Y_n$  is the ground truth of pedestrian *n*, *k* is a hyperparameter, and  $\hat{Y}_n^t$  is the predicted trajectory returned by our network. This loss guides our model to generate the outputs which are consistent with the observed information.

In summary, our method has three context-driven modules for modeling the interactions between human and human, human and global scene, and human and local context. The global scenes and local contexts are represented by the corresponding features extracted by pretrained CNNs. The spatial and temporal correlations based on GATs and LSTMs are used throughout the method. Specifically, we use GAT networks to learn the weights for the human-human and local human-scene interactions among all the pedestrians in the scene and LSTMs to model the temporal correlation of these interactions, as shown in Figs. 4 and 8. For global human-scene interaction, it is processed differently. Although the background scene is almost the same in most cases, there are still some background motions. In order to model these human-scene interactions, a GAT network is used to assign attention weights to the scene features at different time steps, as shown in Fig. 7.

# **4** Experimental results

In this section, we compare the performance of our MCDIM method against the state of the art on two public and commonly used datasets: the walking pedestrians dataset provided by ETH Zurich (ETH) [39] and the crowd data provided by the University of Cyprus (UCY) [40]. In order to evaluate the generalization ability of our method, we also conduct experiments on another two datasets, Town Centre [41] and Grand Central Station [42].

#### 4.1 Datasets

Dataset ETH contains two videos (ETH and HOTEL), and UCY contains three video sequences (UNIV, ZARA1, and ZARA2). These sequences are recorded in 25 frames per second (fps), consisting of 2D real-world human trajectories and bird's-eye view images of four different scene backgrounds. There are 1536 pedestrians in the crowded scenes, which contain a wide variety of human movement patterns, including challenging situations such as pedestrians walking in the same direction, crossing each other, and collision avoidance.

The Town Centre dataset [41] is originally collected to evaluate the performance of person tracking. It contains hundreds of pedestrians in a real-world crowded scene. The annotation file provides bounding boxes of each pedestrian's body and head. Following previous methods [23, 24], we define the location of a pedestrian as the center position of his/her body bounding box. The trajectory data is collected for every five frames. The Grand Central Station dataset [42] is originally collected to analyze human behaviors. It is a long video (about 33:20 minutes) recorded in a crowded station, which contains trajectories of about 12,600 pedestrians. Both Town Centre and Grand Central Station datasets consist of large amounts of human–human and human–scene interactions.

#### 4.2 Implementation details

Our model is constructed using LSTMs and GATs. The hidden state dimensions of the LSTM encoder, the temporal LSTM, and the LSTM decoder are set to 32. Following prior work [18], the scene feature of size 512 is extracted by the VGGNet-19 network [38] which is pretrained on the ImageNet [38]. The local image patch size is  $128 \times 128$ . Two graph attention layers [43] are used to learn the interactions. The hidden state dimension and output dimension of the graph attention layer are 16 and 32, respectively. Our model is trained with the Adam optimizer with an initial learning rate of 0.001 and a batch size of 64. The hyper-parameter *k* in (21) is set as 20.

#### 4.3 Evaluation metrics and protocol

We evaluate our performance using two standard metrics, average displacement error (ADE) and final displacement error (FDE), as in existing methods [14, 17, 18, 20]. ADE defined in (22) computes the average Euclidean distance between the predicted trajectories and the ground truth trajectories from time step  $T_O + 1$  to  $T_P$ . FDE defined in (23) computes the Euclidean distance between the final position of the predicted trajectory and ground truth trajectory at time step  $T_P$ . ADE and FDE are defined as follows:

ADE = 
$$\frac{\sum_{n \in \Omega} \sum_{t=T_{O}+1}^{T_{P}} \sqrt{(\hat{x}_{n}^{t} - x_{n}^{t})^{2} + (\hat{y}_{n}^{t} - y_{n}^{t})^{2}}}{|\Omega| \cdot T_{P}}, \quad (22)$$

$$FDE = \frac{\sum_{n \in \Omega} \sqrt{(\hat{x}_n^{T_P} - x_n^{T_P})^2 + (\hat{y}_n^{T_P} - y_n^{T_P})^2}}{|\Omega|},$$
(23)

where  $(\hat{x}_n^t, \hat{y}_n^t)$  and  $(x_n^t, y_n^t)$  are the predicted and ground truth trajectory coordinates for pedestrian *n* at time *t*,  $\Omega$  represents the set of observed pedestrians and  $|\Omega|$  denotes the total number of pedestrians.

We follow the standard leave-one-out protocol [17] for performance evaluation with 4 datasets being used for training and the remaining one for testing. During the trajectory prediction, the number of time steps of the observed trajectories is 8 (or 3.2 seconds), and the prediction window size is 12 time steps (or 4.8 seconds). For the generalization experiments on the Town Centre and Grand Central Station datasets, we follow the previous work [23] to normalize the location coordinates to [0, 1] and split the whole data into two halves for training and testing.

#### 4.4 Baseline methods

We compare our method against the following state-of-theart methods: (1) S-GAN [17]: This is one of the first GANbased methods. It has two variants, S-GAN and S-GAN-P. which are different in whether applying the pooling mechanism. The hyper-parameter k in the variety loss is set to 20 for evaluation. (2) Sophie [18]: This method observes human trajectories and scene images and applies a soft attention mechanism for both features. (3) Scene-LSTM [23]: This method designs two coupled LSTMs to encode both pedestrian's past trajectories and the scene grids. (4) Next [44]: This method extracts multiple visual features, such as the pedestrian's bounding box, pedestrian keypoints, and semantic scene features for encoding, and uses an LSTM decoder to predict future trajectories. To encourage the multiple feasible paths, the authors train 20 different models with random initialization by following [17]. (5) STGAT [20]: This method captures the spatial interactions by GAT and also designs an extra set of LSTM to extract the temporal information along the observed time steps. (6) Social-BiGAT [21]: This method uses the GAT to model the human-human interactions and applies a soft attention mechanism to the extracted visual features. (7) Reciprocal-GAN [25]: This method constructs a forward network and a backward network based on a reciprocal consistency constraint. (8) SRA-LSTM [15]: This method designs an LSTM as a social relationship encoder to model the temporal correlation of the relative position among pedestrians. (9) AC-VRNN [28]: This method proposes a generative architecture based on Conditional Variational Recurrent Neural Network, which relies on prior belief maps to force the model to consider past observed dynamics in generating future trajectories.

#### 4.5 Quantitative results

The ADE and FDE performance comparison results of the above methods on the ETH and UCY datasets are summarized in Table 1. It can be seen that our method outperforms existing methods in most cases, except on the ETH dataset against the *Scene-LSTM* and ZARA1 dataset against *Reciprocal-GAN*. *S-GAN* and *S-GAN-P* perform the worst as expected since both methods do not utilize the scene information and directly model the human-human interactions based on their distance. Both *Scene-LSTM* and *Sophie* perform better than *S-GAN* due to the use of the scene information. *Next* encodes various visual features extracted from the scene and employs the focal attention on the encoded features, which helps it achieve a better result

20111

Table 1 Comparison results of all the baselines and our method	Metric	Method	ETH	HOTEL	UNIV	ZARA1	ZARA2	Avg
on ETH (Columns 3 and 4) and UCY (Columns 5-7) datasets	ADE	S-GAN, CVPR'18 [17]	0.81	0.72	0.60	0.34	0.42	0.58
		S-GAN-P, CVPR'18 [17]	0.87	0.67	0.76	0.35	0.42	0.61
		Sophie, CVPR'19 [18]	0.70	0.76	0.54	0.30	0.38	0.54
		Scene-LSTM, ISCV'19 [23]	0.36	0.95	0.63	0.45	0.40	0.56
		Next, CVPR'19 [44]	0.73	0.30	0.60	0.38	0.31	0.46
		STGAT, ICCV'19 [20]	0.65	0.64	0.52	0.34	0.29	0.43
		Social-BiGAT, NIPS'19 [21]	0.69	0.49	0.55	0.30	0.36	0.48
		Reciprocal-GAN, CVPR'20 [25]	0.69	0.43	0.53	0.28	0.28	0.44
		SRA-LSTM, 2021 [15]	0.59	0.29	0.55	0.37	0.43	0.45
		AC-VRNN, 2021 [28]	0.61	0.30	0.58	0.34	0.28	0.42
		MCDIM (Our Work)	0.59	0.29	0.51	0.32	0.28	0.40
	FDE	S-GAN, CVPR'18 [17]	1.52	1.61	1.26	0.69	0.84	1.18
		S-GAN-P, CVPR'18 [17]	1.62	1.37	1.52	0.68	0.84	1.21
		Sophie, CVPR'19 [18]	1.43	1.67	1.24	0.63	0.78	1.15
		Scene-LSTM, ISCV'19 [23]	0.67	1.77	1.41	1.00	0.90	1.15
		Next, CVPR'19 [44]	1.65	0.59	1.27	0.81	0.68	1.00
		STGAT, ICCV'19 [20]	1.12	0.66	1.10	0.69	0.60	0.83
		Social-BiGAT, NIPS'19 [21]	1.29	1.01	1.32	0.62	0.75	1.00
		Reciprocal-GAN, CVPR'20 [25]	1.24	0.87	1.17	0.61	0.59	0.90
		SRA-LSTM, 2021 [15]	1.16	0.56	1.19	0.82	0.93	0.93
		AC-VRNN, 2021 [28]	1.09	0.55	1.22	0.68	0.59	0.83
		MCDIM (Our Work)	1.07	0.57	1.09	0.68	0.59	0.80

Given the eight previous time steps, the task predicts 12 future time steps. All methods take ADE / FDE in meter scale as error metrics

than Sophie. Background image features are also encoded in Reciprocal-GAN. Although SRA-LSTM does not take advantage of the physical scene features, it introduces a social relationship attention module to capture the social interactions, which makes a significant contribution to its performance. Similar to SRA-LSTM, STGAT only takes trajectories as input, but it captures the spatial-temporal interactions among pedestrians using a complex GATbased network. Its average performance is 0.43 and 0.83 for ADE and FDE, respectively, the third best among all these methods. AC-VRNN models the human-human interactions using a Conditional Variational Recurrent Neural Network and uses prior belief maps to force the model to consider past observed dynamics in generating future trajectories. It achieves the second best average performance, namely, 0.42 and 0.83 for ADE and FDE, respectively. Our method models the human-human and human-scene interactions at different spatial and temporal scales, in which the impact of global and local scene context is fully considered using the scene context encoders, and the spatial-temporal correlations of the interactions are captured using LSTMs and GATs. The overall performance of our method is the best, which is 0.40 and 0.80 for average ADE and FDE scores, respectively.

We also evaluate another performance metric called near-collisions rate proposed in [18] to evaluate the ability of our method in predicting reasonable and feasible paths in crowded scenes. It represents the probability of two persons moving closer than 0.1m. The average probability (in percentage) of near-collisions across all frames in ETH and UCY datasets are reported in Table 2. The results of S-GAN and Sophie are cited from [18]. We can see that our method outperforms both S-GAN and Sophie, which indicates that our method is able to generate more reasonable paths to prevent collisions.

Table 2 Average percentage of colliding pedestrians for each scene in ETH and UCY datasets

	GT	S-GAN [17]	Sophie [18]	Ours
ETH	0.000	2.509	1.757	1.447
HOTEL	0.092	1.752	1.936	1.326
UNIV	0.124	0.559	0.621	0.514
ZARA1	0.000	1.749	1.094	1.172
ZARA2	0.732	2.020	1.464	1.315
Avg	0.189	1.717	1.361	1.155

The first column represents the ground truth

# 4.6 Generalization capability evaluation on the town centre and grand central station datasets

We conduct experiments on two new datasets, the Town Centre [41] and Grand Central Station [42], to further evaluate the generalization capability of our method. The experiment settings are exactly the same as [23]. For both datasets, the training data is formed by combining the training data from ETH and UCY datasets and 50% data from this dataset. The remaining data is used for testing. We compare the results of trajectory prediction for prediction window sizes of 12 and 16 time steps against the *S*-*GAN* [17] and *Scene-LSTM* [23] methods. As illustrated in Table 3, our method clearly outperforms these two state-of-the-art methods. Some qualitative examples from both datasets are shown in Fig. 9.

# 4.7 Qualitative results

As mentioned before, human trajectory prediction is a challenging task, and our goal is to predict both socially and physically acceptable future paths. Many complicated interactions may occur between pedestrians in a crowded scene, such as group movements, walking toward each other, and changing directions to avoid collisions with other pedestrians or stationary obstacles.

Figure 10 shows some qualitative results of future trajectory prediction with different human-human and human-scene interactions. For example, in the first figure of the first row, a pedestrian walks out of the building, and three other pedestrians walk in the same direction without collisions. In the second figure of the first row and the last figure of the second row, our model can learn the interactions between the pedestrian and the obstacles (*e.g.*, cars and trees), therefore changing the trajectories' directions. From the last figure of the first row and the first figure of the second row, we can see that our model can generate reasonable and feasible paths for multiple pedestrians walking toward each other. Our model also performs well for the pedestrians that are not moving. As shown in the second and third figures of the second row, our model can learn the future trajectories of the pedestrians standing in the same place waiting for the train or chatting with each other. These examples demonstrate that our method can predict reasonable and feasible future trajectories in complex and crowded scenes.

# 4.8 Ablation studies

Compared with the existing methods, our method has two major new components, the global scene layout encoder, and the local scene context encoder. In order to evaluate the importance and contribution of each new component, several ablation experiments are performed with the ADE and FDE results reported in Table 4. From the second and fifth rows of Table 4, we can see that the error metrics ADE and FDE both increase when we exclude the global scene layout encoder. With the global scene layout encoder, our model can model and learn the impact of the entire scene layout on the human trajectory to generate better and more reasonable future trajectories. In the third and sixth rows of Table 4, we list the results without the local scene context encoder. Both ADE and FDE scores increase since the model lacks the knowledge of the scene context in the local neighborhood. From the last column of Table 4, we can see that our two new components have significantly contributed to the overall performance.

# 4.9 Discussion and future work

In the previous section, we have compared our method to the state of the art. In Table 1, we have the following observations on the factors that would improve the

Metrics	Datasets	Prediction Length	S-GAN [17]	S-GAN-P [17]	Scene-LSTM [23]	Ours
ADE	Town center	12	0.22	0.21	0.09	0.07
		16	0.37	0.38	0.14	0.10
	Grand central station	12	0.21	0.40	0.11	0.08
		16	0.32	0.79	0.14	0.11
FDE	Town center	12	0.46	0.42	0.18	0.11
		16	0.80	0.81	0.27	0.19
	Grand central sation	12	0.45	0.74	0.17	0.14
		16	0.62	1.50	0.25	0.17

Table 3 The quantitative results (ADE and FDE) on town centre and grand central station datasets with different prediction lengths of future trajectories



Fig. 9 Qualitative examples of our method predicting future 12 time steps trajectories, given previous 8 time steps ones on town centre (1st row) and grand central station (2nd row) dataset. Note that we crop and resize the original image for better visualization



**Fig. 10** Qualitative examples of our method predicting future 12 time steps trajectories, given previous 8 time steps ones on ETH and UCY dataset. Note that we crop and resize the original image for better

**Table 4** Ablation experimentsof our full algorithm withoutdifferent components

visualization. We choose scenarios with multiple pedestrians and complex interactions

Metric	Method	ETH	HOTEL	UNIV	ZARA1	ZARA2	Avg
ADE	Our method (full algorithm)	0.59	0.29	0.51	0.32	0.28	0.40
	Without global scene layout encoder	0.66	0.47	0.52	0.30	0.30	0.45
	Without local scene context encoder	0.63	0.45	0.52	0.31	0.29	0.44
FDE	Our method (full algorithm)	1.07	0.57	1.09	0.68	0.59	0.80
	Without global scene layout encoder	1.25	0.72	1.10	0.69	0.60	0.87
	Without local scene context encoder	1.28	0.78	1.11	0.68	0.61	0.89

Error metrics reported are ADE and FDE in meter scale

trajectory prediction performance. The first one is incorporating the scene features into the interaction modeling scheme, as shown in the performance comparison between *Sophie*, *S-GAN*, and ours. The second one is modeling the spatial-temporal correlations of human interactions, as demonstrated by the improvements achieved by *STGAT*. Our method combines the two and models the interactions between human and human, interactions between human and global and local scenes. The spatial-temporal correlations of these interactions are also captured using LSTMand GAT-based networks. These factors help our method achieve the best average performance. The ablation study in Table 4 shows that both global and local human–scene interactions have made significant contributions to the final performance.

On the other hand, in Table 1, our method fails to achieve the best for some tests. For example, *Scene-LSTM* and *Reciprocal-GAN* have better performance than ours on ETH and ZARA1, respectively. In these two video sequences, the background scenes are quite stationary, especially the ETH. Therefore, the background image features are almost the same for trajectory prediction at different time steps. This suggests it benefits to design a model to better understand the scene, such as utilizing the semantic segmentation of the images, as the work of [32]. Furthermore, the comparison among those methods that only utilize the trajectory information indicates that applying a more advanced learning network or algorithm will certainly improve the performance. Compared to the network architectures of *SRA-LSTM* and *AC-VRNN*, our network structure is quite straightforward, which may limit its learning capacities. In future work, we can further improve our method by designing a more advanced network that is more capable of capturing the complex hidden correlations and interactions among pedestrians and scenes. We also plan to test the human trajectory prediction in more crowded and dynamic scenarios, such as players in sports games.

# 5 Conclusion

In this work, we have recognized that the human-human interactions and the interactions between humans and the multi-level scene contexts occur at different spatial and temporal scales in real scenarios, which have not been fully addressed for human trajectory prediction. To capture these complex interactions, we propose a multi-level contextdriven interaction modeling (MCDIM) method for human future trajectory learning and prediction. We construct a multilayer GAT network to model the hierarchical humanhuman interactions. An extra set of LSTMs has been designed to capture the correlations of these human-human interactions at different temporal scales. Both the global scene layout features and local scene context features in the neighborhood of the person at different time steps are extracted and encoded with LSTMs. The hidden states are fed into the GAT network for joint learning of humanhuman and human-scene interactions. Experimental results on several benchmark datasets demonstrate that, by jointly modeling the human-human and human-scene interactions at different spatial-temporal scales, our method outperforms the state-of-the-art methods and generates more accurate and plausible trajectories for pedestrians.

**Funding** This work was supported by the National Natural Science Foundation of China under grant 61971290, and the Shenzhen Stability Support General Project (Category A) 20200826104014001.

#### Declarations

**Conflict of Interest** The authors declare that they have no conflict of interest.

**Ethical approval** This article does not contain any studies with human participants or animals performed by any of the authors.

**Informed consent** Informed consent was obtained from all individual participants included in the study.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

# References

- Srikanth S, Ansari JA, Sharma S et al. (2019) Infer: Intermediate representations for future prediction. arXiv preprint arXiv:1903. 10641
- 2. Abel Díaz Berenguer (2020) Mitchel Alioscha-Perez, Meshia Cédric Oveneke, and Hichem Sahli. In: Context-aware human trajectories prediction via latent variational model, IEEE Transactions on Circuits and Systems for Video Technology
- XT Truong, N Trung Dung (2017) Toward socially aware robot navigation in dynamic and crowded environments: A proactive social motion model. IEEE Trans Autom Scie Eng 14(4):1743–1760
- Ji Y, Yang Y, Shen F, Heng TS, Xuelong L (2019) A survey of human action analysis in hri applications, In: IEEE Transactions on Circuits and Systems for Video Technology
- Wang Z, Liu S, Zhang J, Chen S, Guan Q (2016) A spatiotemporal crf for human interaction understanding. IEEE Trans Circuits Syst Video Technol 27(8):1647–1660
- Alejandro B, Pietro M, Regazzoni CS, Matthias R (2015) The evolution of first person vision methods: a survey. IEEE Trans Circuits Syst Video Technol 25(5):744–760
- Luber M, Stork JA, Tipaldi GD, Arras KO. (2010) People tracking with human motion predictions from social forces. In: 2010 IEEE International Conference on Robotics and Automation, pp 464–469. IEEE
- Mehran R, Oyama A, Shah M (2009) Abnormal crowd behavior detection using social force model. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp 935–942. IEEE
- Azhar A, Rubab S, Khan M, Bangash YA, Alshehri MD, Illahi F, Bashir AK (2022) Detection and prediction of traffic accidents using deep learning techniques. Cluster Comput, 1–17
- Mohamed A, Qian K, Elhoseiny M, Claudel C (2020) Socialstgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 14424–14432
- Anirudh V, Katharina M, Jean O (2018) Social attention: modeling attention in human crowds. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp 1–7. IEEE
- Pellegrini S, Ess A, Schindler K, Van Gool L (2009) You'll never walk alone: Modeling social behavior for multi-target tracking. In: 2009 IEEE 12th International Conference on Computer Vision, pp 261–268. IEEE
- 13. Trautman P, Krause A (2010) Unfreezing the robot: Navigation in dense, interacting crowds. In: 2010 IEEE/RSJ International

Conference on Intelligent Robots and Systems, pp 797–803. IEEE

- 14. Alahi A, Goel K, Ramanathan V, Robicquet A, Fei-Fei L, Savarese S (2016) Social lstm: human trajectory prediction in crowded spaces. In: Proceedings of the IEEE conference on computer vision and pattern recognition,pp 961–971
- Peng Y, Zhang G, Shi J, Xu B, Zheng L (2021) SRA-LSTM: social relationship attention LSTM for human trajectory prediction. *CoRR*, abs/2103.17045
- Spp Hochreiter, Jürgen Schmidhuber (1997) Long short-term memory. Neural Comput 9(8):1735–1780
- Gupta A, Johnson J, Fei-Fei L, Savarese S, Alahi A (2018) Social gan: Socially acceptable trajectories with generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2255–2264
- Sadeghian A, Kosaraju V, Sadeghian A, Hirose N, Rezatofighi H, Savarese S (2019) Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1349–1358
- Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680
- Huang Y, Bi H, Li Z, Mao T, Wang Z (2019) Stgat: modeling spatial-temporal interactions for human trajectory prediction. In: Proceedings of the IEEE International Conference on Computer Vision, pp 6272–6281
- Kosaraju V, Sadeghian A, Martín-Martín Roberto, Reid Ian, Rezatofighi Hamid, Savarese Silvio (2019) Social-bigat: multimodal trajectory forecasting using bicycle-gan and graph attention networks. In: Advances in Neural Information Processing Systems, pp 137–146
- Veličković Petar, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y (2017) Graph attention networks. arXiv preprint arXiv: 1710.10903
- Huynh M, Alaghband G (2019) Trajectory prediction by coupling scene-lstm with human movement lstm. In: International Symposium on Visual Computing, pp 244–259. Springer
- 24. Xue H, Huynh DQ, Reynolds M (2018) Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp 1186–1194. IEEE
- Sun H, Zhao Z, He Z (2020) Reciprocal learning networks for human trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7416–7425
- Helbing D, Molnar P (1995) Social force model for pedestrian dynamics. Phys Rev E 51(5):4282
- Bisagno Niccoló, Zhang B, Conci N (2018) Group Istm: Group trajectory prediction in crowded scenarios. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0
- Bertugli Alessia, Calderara Simone, Coscia Pasquale, Ballan Lamberto, Cucchiara Rita (2021) AC-VRNN: attentive conditional-vrnn for multi-future trajectory prediction. Comput Vis Image Underst 210:103245
- 29. Zhang J, Zhenqiang M, Guo Y, Obaidat MS (2021) ATPS: an adaptive trajectory prediction system based on semantic

information for dynamic objects. Neural Comput Appl 33(17):11447–11464

- Liang J, Jiang Lu, Murphy K, Yu T, Hauptmann A (2019) The garden of forking paths: towards multi-future trajectory prediction. arXiv preprint arXiv:1912.06445
- 31. Manh H, Alaghband G (2018) Scene-lstm: A model for human trajectory prediction. arXiv preprint arXiv:1808.04018
- 32. Lisotto M, Coscia P, Ballan L (2019) Social and scene-aware trajectory prediction in crowded spaces. In: Proceedings of the IEEE International Conference on Computer Vision Workshops
- ibitemkipf2016semi Kipf Thomas N, Welling Max (2016) Semisupervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907
- 34. Si C, Chen W, Wang W, Wang L, Tan T (2019) An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1227–1236
- 35. Yan S, Xiong Y, Lin D (2018) Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirtysecond AAAI Conference on Artificial Intelligence
- 36. Hastie T, Tibshirani R (2009) The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media
- Talathi Sachin S, Vartak A (2015) Improving performance of recurrent neural network with relu nonlinearity. arXiv preprint arXiv:1511.03771
- Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis 115(3):211–252. https://doi.org/10.1007/ s11263-015-0816-y
- Pellegrini S, Ess A, Van Gool L (2010) Improving data association by joint modeling of pedestrian trajectories and groupings. In: European Conference on Computer Vision, pp 452–465. Springer
- 40. Leal-Taixé L, Fenzi M, Kuznetsova A, Rosenhahn B, Savarese S (2014) Learning an image-based motion context for multiple people tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3542–3549
- 41. Benfold B, Reid I (2011) Stable multi-target tracking in real-time surveillance video. In CVPR 2011, pp 3457–3464. IEEE
- 42. Zhou B, Wang X, Tang X (2012) Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrianagents. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp 2871–2878. IEEE
- 43. Qiu J, Tang J, Ma H, Dong Y, Wang K, Tang J (2018) Deepinf: Social influence prediction with deep learning. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'18)
- 44. Liang J, Jiang L, Niebles Juan C, Hauptmann Alexander G, Fei-Fei L (2019) Peeking into the future: Predicting future person activities and locations in videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5725–5734

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.