Recursive Social Behavior Graph for Trajectory Prediction

Jianhua Sun[1], Qinhong Jiang[2], Cewu Lu[1†]
[1] Shanghai Jiao Tong University, China
[2] SenseTime Group Limited, China
{gothic, lucewu}@sjtu.edu.cn jiangqinhong@sensetime.com

Abstract

Social interaction is an important topic in human trajectory prediction to generate plausible paths. In this paper, we present a novel insight of group-based social interaction model to explore relationships among pedestrians. We recursively extract social representations supervised by group-based annotations and formulate them into a social behavior graph, called Recursive Social Behavior Graph. Our recursive mechanism explores the representation power largely. Graph Convolutional Neural Network then is used to propagate social interaction information in such a graph. With the guidance of Recursive Social Behavior Graph, we surpass state-of-the-art method on ETH and UCY dataset for 11.1% in ADE and 10.8% in FDE in average, and successfully predict complex social behaviors.

1. Introduction

Forecasting the future trajectory of humans in a dynamic scene is an important task in computer vision[28, 16, 31, 32, 33, 42, 44, 20]. It is also one of the key points in autonomous driving and human-robot interaction, which explores dense information for the following decision making process. A main challenge of trajectory forecasting lies in how to incorporate human-human interaction into consideration to generate plausible paths [2, 13, 3, 6, 27, 26].

Early works have made a lot effort to solve the problem. Social Force [14, 28] abstracts out different types of force, such as acceleration and deceleration forces to handle it. In recent years, great progress has been made in deep learning, which inspired researches start working on Deep Neural Networks based methods. Some researches [2, 13, 34, 18, 17] modified Recurrent Neural Networks (RNNs) architecture with particular pooling or attention mechanism to integrate information between RNNs.

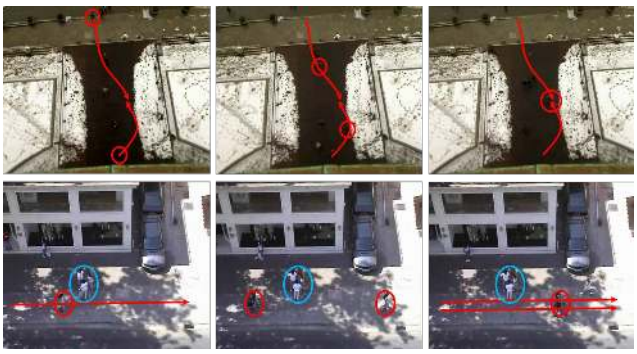

Figure 1. Examples of distant unrelated human-human interactions. Images are in chronological order from left to right. The top three images show that two people (with red circle) walk to the same destination from opposite directions. The bottom three images show people with left red circle are following the person in right red circle with little impact from people in blue circle.

Although great improvements have been made, there still exists challenges. Force based models[28] utilize the distance to compute force, and will fail when the interaction is complicated. And for pooling methods [2, 13], the distance between two person at a single timestep is used as a criterion to calculate the strength of the relationship. Attention method in [18, 34] also meet the same problem that Euclidean distance are used in their method to guide the attention mechanism. In general, these learning methods try to use distance to formulate the strength of influences between different agents, but ignore that distance-based scheme cannot handle numerous social behaviours in human society. Fig. 1 shows two typical examples. The top three images show that two people walk to the same destination from opposite directions. The bottom three images show three pedestrians walk along the street while another three person stand still and talk with each other. Even though pedestrians in red circles in these two scenes are in a great distance, they show a strong relationship.

In this paper, we aim to explore relationships among pedestrians beyond the use of distance. To this end, we present a new insight of **group-based** social interaction modeling. A group can be defined as a set of people with

†Cewu Lu is corresponding author, member of Qing Yuan Research Institute and MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China.

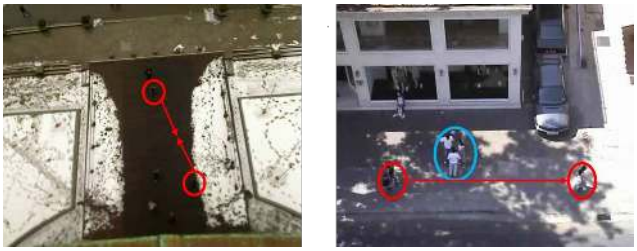

Figure 2. Examples of groups and interaction in groups. Red and blue circles are different groups. The direction of arrow represents the direction of influence in interaction.

similar movements, behaviours, purpose or destinations. As shown in Fig. 2, each color represents a group and the relations are annotated with arrows to show the directionality of interactions. Further, such groups in a scene can be formulated as a graph, which is a common structure for feature propagation. Additionally, we argue that social relationship representation is too complicated and cannot well be captured by hand-crafted methods.

To model this novel insight, we present a neural network to recursively extract social relationships and formulated them into a social behavior graph, called Recursive Social Behavior Graph (RSBG). Each pedestrian is considered as a node with features that takes historical trajectories into consideration. Those nodes are connected by relational social representations which are considered as the edges of the graph. We use group annotations to supervise the generation of social representation, which is the first time social related annotations are used to help neural networks to learn social relationships as far as we know. Moreover, a recursive mechanism is introduced. We recursively update individual trajectory features in interaction scope by social representations, in turn, better individual features are used to upgrade social representations. To propagate features guided by RSBG, our system works under a framework of Graph Convolutional Neural networks (GCNs).

Experiments on multiple human trajectory benchmark, including two datasets in ETH[31] and three datasets in UCY[21], show the superior of our model in accuracy improvement. Our contributions can be summarized as follows:

1. We propose Recursive Social Behavior Graph, a novel graph representation for social behaviour modeling, and a recursive neural network to generate it. The network is designed to extract latent pedestrian relationships and is supervised by group annotations, which is the first time that social related annotations annotated by experts are introduced in prediction tasks.

2. We first introduce GCNs to integrate human social behaviours in dynamic scenes for prediction task, which leads to greater expressive power and higher performance.

3. We conduct exhaustive experiments in several video datasets. By applying our proposed approach, we are able to achieve 11.1% improvement in ADE and 10.8% in FDE comparing with state-of-the-art method.

2. Related Work

Human trajectory forecasting. Human trajectory forecasting is a task to predict possible trajectories of a person according to his historical trajectory and vision based features, such as his current actions and surroundings. With the maturity of human understanding and trajectory tracking techniques[30, 6, 10, 12, 11], numerous studies has been done in this field[28, 16, 31, 32, 33, 42, 44, 20, 5]. Early researches [28, 39, 20] try to build mathematical models to predict the trajectory. For example, Energy Minimization[39] model constructs a grid graph with costs on each edges, formulates trajectory prediction as a shortest path problem and solves it by Dijkstra algorithm. IRL proposed by Abbeel *et al*. [1] has been used for trajectory prediction in [20], which models human behaviour as a sequential decision-making process.

With the development of neural networks, many prediction methods [22, 2, 13, 17, 34, 35, 41] based on deep learning has been proposed, and focused on different insights to solve this problem. Alahi *et al*. [2] modified vanilla LSTM structure using a novel pooling mehtod to propagate human interactions in crowd scenes. Gupta *et al*. [13] and Li *et al*. [22] applied a Generative Adversarial Network in their prediction framework to explore the multimodality of human behaviours. Sadeghian *et al*. [34] and Liang *et al*. [25] extracted rich information from context for more accurate predictions. All these researches have made a huge breakthrough.

Human-human interactions in trajectory forecasting. Human object interaction (HOI) [9, 36, 24, 40, 23] brings abundant information for scene understanding. Thus, human-human interaction is critical to predict future trajectories correctly. Early researches, such as Social Forces[14], modeling human-human interactions in dynamic scenes by various types of forces. However, as some key parameters are highly based on prior knowledge, such as force definition, they cannot handle sophisticated and crowd scene with all kinds of pedestrians who may act totally different.

Recent years, Recurrent Neural Network (RNN) has shown great power for sequence problems[4, 7, 8, 29, 19]. However, single RNN based architecture cannot deal with human-human interaction. Alahi *et al*. [2] proposed Social-LSTM which applies social pooling after each time step in vanilla LSTM to integrate social features. Gupta *et al*. [13] improved social pooling to capture global context. These

pooling methods use distance between two person as a criterion to calculate the strength of the relationship. Further, [34, 18] introduced attention mechanism to propagate social features, but they also meet the problem that the attention are highly restricted by distance. Sadeghian *et al*. [34] using Euclidean distance between target agent and other agents as a reference to permute these agents for permutation invariant before attention mechanism, while Ivanovic *et al*. [18] using Euclidean distance to build a traffic agent graph to guide attention mechanism. Thus these methods cannot handle the situations described in Fig. 1 very well.

Recently, Huang *et al*. [17] proposed a Graph Attention (GAT) based network to propagate spatial and temporal interactions between different pedestrians without particular supervision for attention mechanism. Although this method is not restricted by distance, but the attention mechanism cannot handle sophisticated scenes because of the lack of supervision and may fail in certain cases as discussed in Sec. 4.2 in [17].

Graph Neural Network. Graph Neural Network (a.k.a. GNN) and its variants[38] are born to handle data represented in the Euclidean space. GNNs can be categorized into different types, and among them Graph Convolutional Networks (GCNs)[15] have been widely used in different computer vision tasks. For instance, Gao *et al*. [12] trains GCNs in a deep siamese network for robust target localization in visual tracking task. STGCN, a variant of normal GCN, is used by Yan *et al*. [43] to build a dynamic skeleton graph for human action recognition. Wang *et al*. [37] adopts GCN to match graphs in images. In this paper, we will show how GCNs propagate social features during human-human interaction and successfully improve overall accuracy on trajectory prediction.

3. Approach

In this section, we propose a social behavior driven model to enable trajectory prediction from group level. It is designed to capture the fact that pedestrians in public places often gather and walk in groups, especially in crowd scenes. These groups apparently demonstrate remarkable social behaviors, such as following and joining, which is important for trajectory prediction.

3.1. Problem Definition

Following previous works [2, 13], we assume that each video is preprocessed by detection and tracking algorithm to obtain the spatial coordinates and specific ID for each person at each timestep. Therefore, at a certain timestep t for person ID i, we can formulate his/her coordinate as (x_i^t, y_i^t), and the frame-level surrounding information as S_i^t, e.g. a top-view or angleview image patch centered on person i at time t. We observe the coordinate sequences and the instance patch for everyone in time step $[1, T_{obs}]$ as input, and forecast the coordinate sequences in $[T_{obs+1}, T_{obs+pred}]$ as output.

3.2. Overview

Given a series of pedestrians together in a scene provided by a video, the relationship between each pair of them can be defined by a set

$$\mathbf{R} = \{\mathbf{r}(\mathbf{i_1}, \mathbf{i_2}) | \mathbf{0} \leq \mathbf{i_1}, \mathbf{i_2} < \mathbf{N}, \mathbf{i_1} \neq \mathbf{i_2}\} \quad (1)$$

where i_x denotes the unique ID for each person in the scene, N denotes the total number of pedestrians in the scene, and $r(i, j)$ denotes the relational social representation between the i_{th} and j_{th} person. With individual representation for each person $\mathbf{f_i}$, the relationship set can be formulated into a social behavior graph $\mathcal{G}$. We design the individuate representation and relational social representation as node and edge features of $\mathcal{G}$ respectively. Thus, a novel recursive framework is preformed on $\mathcal{G}$ to better understand social relationship, we call it as Recursive Social Behavior Graph (RSBG). Given the powerful feature from recursive $\mathcal{G}$, we can predict future trajectory by LSTM model.

In the following sections, we will introduce individual representation in Sec. 3.3 and relational social representation in Sec. 3.4. The recursive social behavior graph (RSBG) will be discussed in Sec. 3.5. Finally, in Sec. 3.6, we introduce how to integrate proposed RSBG into the LSTM for high quality trajectory prediction.

3.3. Individual Representation

We adopt historical trajectory feature and human context feature as our individual representation.

Historical Trajectory feature In real social dynamic scenes, people will act after deciding the path in several seconds as a general rule, which means later trajectories will largely influence the former ones. By this end, we adopt a BiLSTM architecture instead of popular vanilla LSTM [2, 13] to capture individual feature, considering the dependencies of both previous and future steps, which could generate a more comprehensive representation for individual trajectory.

Human Context feature To extract frame-level human instance context information, we use Convolutional Neural Network (CNN). Specifically, for each spatial position (x_i^t, y_i^t) at timestep t of pedestrian i, we can obtain a image patch s_i^t from video centered on (x_i^t, y_i^t). Therefore, for a whole historical trajectory of person i, we feed the patch set $\mathcal{S}_i = \{s_i^t, 0 \leq t < T_{obs}\}$ into the CNN framework to calculate visual information $\mathbf{V_i}$, which can be represented as human context feature.

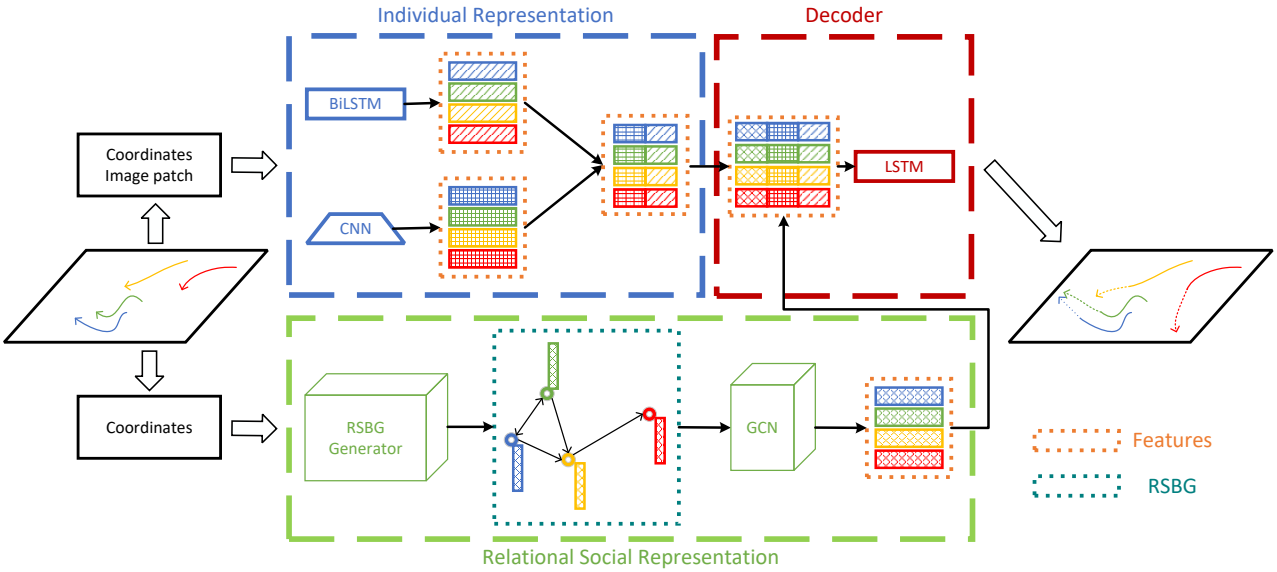

Figure 3. Overview of our proposed prediction method. For individual representation, BiLSTMs are used to encode historical trajectory feature, and CNNs are used to encode human context feature. For relational social representation, we first generate RSBG recursively and then use GCN to propagate social features. At the decoding stage, social features are concatenated with individual features which finally decoded by an LSTM based decoder.

Finally, we concatenate historical trajectory feature and context human feature as individual representation. We denote the feature map of i^{th} person instance as $\mathbf{f_i}$.

3.4. Relational Social Representation

Most of the existing social models [2, 13, 18, 34] meets a limitation that they use the distance of pedestrians as a strong reference to build social representation. However, relational social behavior is complicated and can't be easily modeled by a single hand-crafted feature. Therefore, we directly annotate social relationship and learn what is social relationship.

Relationship Labeling In order to supervise training, we introduce social related annotations. In the annotations, pedestrians are separated into groups according to the videos, which can be reconstructed into adjacency matrices, using 0/1 to represent whether two pedestrians are in the same group.

We invite experts who have sociology background to judge relationship of two pedestrians. In the annotation process, experts determine a group of people on the basis of not only physical rules such as velocity, acceleration, direction and relative distance between people, but also sociological knowledge. Considering the group information is dynamic to some extent in a real scene, we split the whole scene into time periods, which is small enough in response to dynamic changes in relationship. Experts annotate the interactions for each time period.

Feature Design For an N people scene, we can construct a feature matrix $\mathbf{F} \in \mathbb{R}^{\mathbf{N}\times\mathbf{L}}$ where each row represents a feature of a certain person, L represents for feature length. Then we define a relation matrix R

$$\mathbf{R} = \mathrm{softmax}(\mathbf{g_s}(\mathbf{F})\mathbf{g_o}(\mathbf{F})^{\mathbf{T}}) \in \mathbb{R}^{\mathbf{N}\times\mathbf{N}} \quad (2)$$

where $g_s(\cdot)$ and $g_o(\cdot)$ are two fully connected layer network function to map $\mathbf{F}$ to two different feature space, we call them subject feature space and object feature space respectively. It is because the pedestrian graph is directional, we need these two functions to guarantee non-commutativity. By integrating subject feature and object feature with an inner product for every ordered pair, an relational embedding matrix $\mathbf{R}$ can response the relationship between any pair of pedestrians. Our relationship labeling provides ground truth of $\mathbf{R}$: 0/1 to represent whether two pedestrians are in the same group.

3.5. Recursive Social Behavior Graph

We design a recursive mechanism to further advance our representations $\mathbf{R}$ and $\mathbf{F}$. First, our individual social representation F should consider the interaction persons around it. Second, we hope the relationship model based on stronger individual social representation. We have the recursive update as

$$\mathbf{R_k} = \mathrm{softmax}(\mathbf{g_s}(\mathbf{F_k})\mathbf{g_o}(\mathbf{F_k})^{\mathbf{T}}) \in \mathbb{R}^{\mathbf{N}\times\mathbf{N}} \quad (3)$$

$$\mathbf{F_{k+1}} = \mathbf{fc}(\mathbf{F_k} + \mathbf{R_k}\mathbf{F_k}) \tag{4}$$

where $\mathbf{fc}$ represents fully connection operation and k is the depth of the recursion. For initialization, features in $\mathbf{F_0}$ are historical trajectories in global coordinate. Formula 4 combines the original information of every person extracted in depth k and interaction information according to groups represented by $\mathbf{R_k}$, which gives an information-rich tensor for the next relational embedding in depth $k+1$.

In our experiments, we set $k = 0, 1, 2$ to extract three relation matrices $(\mathbf{R_0}, \mathbf{R_1}, \mathbf{R_2})$, and fuse them together by $\mathbf{R_a} = \mathbf{Avg}(\mathbf{R_0}, \mathbf{R_1}, \mathbf{R_2})$, where $\mathbf{R_a}$ contains recursive relational features from three stages, and can be viewed as an adjacency matrix for the following graph convolution. We use Cross Entropy Loss here to calculate the loss between ground truth $\mathbf{R}$ and $\mathbf{R_a}$.

With $\mathbf{R_a}$ generated recursively, Recursive Social Behavior Graph (RSBG) is defined as following:

$$\mathcal{G}_{RSB} = (\mathcal{V}, \mathcal{E}) \tag{5}$$

$$\mathcal{V} = \{v_i = \mathbf{t_i} | \mathbf{0} \leq \mathbf{i} < \mathbf{n}\} \tag{6}$$

$$\mathcal{E} = \{e_{i_1 i_2} = \mathbf{R_a}(\mathbf{i_1}, \mathbf{i_2}) | \mathbf{0} \leq \mathbf{i_1}, \mathbf{i_2} < \mathbf{n}, \mathbf{i_1} \neq \mathbf{i_2}\} \tag{7}$$

where $\mathbf{t_i}$ represents the relative historical trajectory for the i_{th} person and $\mathbf{R_a}(\mathbf{i_1}, \mathbf{i_2})$ represents the float in row i_1 column i_2 in $\mathbf{R_a}$. By mapping individual trajectory and relational social representation as vertices and edges respectively, RSBG provides abundant information for following trajectory generation process.

3.6. Trajectory Generation

Graph Convolution Previous works using specially designed pooling method [2, 13] or attention model [18, 34] to propagate social interaction information. In our work, we first introduce Graph Convolutional Network (GCN) to integrate messages guided by RSBG, since GCNs have demonstrated powerful capabilities in processing graph-structured data.

Here, we use GCNs as a message passing scheme to aggregate high-level social information from adjacency nodes, according to $\mathcal{G}_{RSB}$:

$$h_i^m = \frac{\sum_{j \in [0,N), j \in \mathbb{N}} v_j^{m-1} e_{ij}}{\sum_{j \in [0,N), j \in \mathbb{N}} e_{ij}} \tag{8}$$

$$v_i^m = f_{update}(h_i^m) = ReLU(fc(h_i^m)) \tag{9}$$

Formula.8 passes the interaction along weighted edges in $\mathbf{R_a}$. The aggregated features from adjacent nodes are normalized by the total weights of adjacent nodes, as a common practice in GCNs, in order to avoid the bias due to the different numbers of neighbors owned by different nodes. Eq.9 accumulates information to update the state of node i, and f_{update} may take any differentiable mapping function from tensor to tensor. Here, we use a fully connection layer for mapping with $ReLU$ activation. m represents the depth of GCNs and h represents intermediate feature. In our experiments, we use a two-layer GCN network to propagate interaction information which means $m = 1, 2$. Finally, social representation for the i_{th} person can be formulated as $\mathbf{u_i} = \mathbf{v_i^2}$. Note that we use GCN instead of ST-GCN in [43] or GAT in [17] since latent relationship have already fully captured in Relational Social Representation and we only need to propagate features here.

LSTM decoder With previous encoded individual representation features and social representation features, we propose an LSTM based decoder for trajectory generation, where the input $h_i^0 = [\mathbf{f_i}, \mathbf{u_i}]$, and the output is $\hat{Y}_i^t$, representing the coordinate of person id i in timestep t.

Exponential L2 Loss Previous works [13, 25] using L2 loss to evaluate differences between predicted results and ground truth. However, this loss function does not highlight enough on FDE while FDE is a very important indicator to measure prediction accuracy.

By this end, we propose a novel Exponential L2 Loss

$$\mathcal{L}_{EL2}(\hat{Y}_i^t, Y_i^t) = ||\hat{Y}_i^t - Y_i^t||^2 \times e^{\frac{t}{\gamma}} \tag{10}$$

which multiples a coefficient growing over time comparing with L2 loss. Here, $\hat{Y}_i^t$ and Y_i^t are predicted and ground truth coordinate for person i at time t respectively, and γ is a hyper parameter related to T_{pred}. In our experiments, we set it as 20. In Sec. 4.2, we will show Exponential L2 loss gives considerable improvement in FDE metrics and associated improvement in ADE metrics.

4. Experiments

Performance of our models are evaluated on popular benchmarks, including ETH [31] and UCY [21]. ETH and UCY dataset are widely used for human trajectory forecasting benchmark [2, 13, 3, 25, 34]. They contain totally five pedestrian cases in crowd scenes including ETH, HOTEL, UNIV, ZARA1 and ZARA2. We use the same configuration for evaluation following previous work [13]. In detail, we observe trajectories for $3.2sec$ (8 frames) and predict for $4.8sec$ (12 frames) at a frame rate of 0.4, and use a leave-one-out approach for training and evaluation.

Evaluation Metrics. Following previous works [2, 13, 22, 18], we introduce 2 common metrics for testing.

1. *Average Displacement Error* (ADE): Average L2 distance between the ground truth and predicted trajectories.

Method	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
Vanilla LSTM	1.09/2.41	0.86/1.91	0.61/1.31	0.41/0.88	0.52/1.11	0.70/1.52
Social LSTM[2]	1.09/2.35	0.79/1.76	0.67/1.40	0.47/1.00	0.56/1.17	0.72/1.54
Social GAN(1V-1)[13]	1.13/2.21	1.01/2.18	0.60/1.28	0.42/0.91	0.52/1.11	0.74/1.54
PITF[25]	0.88/1.98	0.36/0.74	0.62/1.32	0.42/0.90	0.34/0.75	0.52/1.14
STGAT(1V-1)[17]	0.88/1.66	0.56/1.15	**0.52/1.13**	0.41/0.91	0.31/0.68	0.54/1.11
RSBG w/ context	**0.79/1.47**	0.35/0.71	0.68/1.39	0.42/0.89	0.35/0.71	0.52/1.03
RSBG w/o context	0.80/1.53	**0.33/0.64**	0.59/1.25	**0.40/0.86**	**0.30/0.65**	**0.48/0.99**

Table 1. Comparison with baseline methods on ETH and UCY benchmark for $T_{pred} = 12$ (ADE/FDE). Each row represents a method and each column represents a dataset. 1V-1 means that not use variety loss and sample once during test time according to [13, 17], which simplifies SGAN and STGAT from multimodal to unimodal.

2. *Final Displacement Error* (FDE): The L2 distance between the ground truth destination and the predicted destination at the last prediction timestep.

Benchmarks. We compare with the following baselines, some of them represent state-of-the-art performance in trajectory prediction task.

1. *Vanilla LSTM*: An LSTM network without taking human-human interaction into consideration.
2. *Social LSTM*: Approach in [2]. Each pedestrian is modeled by an LSTM, while hidden states of pedestrians in a certain neighbourhood are pooled at each timestep using Social Pooling.
3. *Social GAN*: Approach in [13]. Each pedestrian is modeled by an LSTM, while hidden states of all pedestrians are pooled at each timestep using Global Pooling. GAN is introduced to generate multimodal prediction results.
4. *PITF*: Approach in [25]. Each pedestrian is modeled by a Person Behavior Module, while person-scene and person-objects interactions are modeled by a Person Interaction Module.
5. *STGAT*: Approach in [17]. Pedestrian motion is modeled by an LSTM, and the temporal correlations of interactions is modeled by an extra LSTM. GAT is introduced to aggregate hidden states of LSTMs to model the spatial interactions.
6. *RSBG*: The method proposed in this paper. We report two different versions of our model: **RSBG w/ context** and **RSBG w/o context**, which represents using and not using human context feature respectively.

Discussion. Some of previous works [13, 34, 17] focused on multimodal prediction (a.k.a. generating multiple trajectories for each single person), which does make sense in real scene. However, as discussed in [18], the BoN evaluation metric in their experiments harms real-world applicability as it is unclear how to achieve such performance online without a prior knowledge of the lowest-error trajectory. Therefore, we mainly focus on unimodal prediction (gives one certain prediction result) to avoid questioning evaluation metric, which means that we test the performance of Social GAN and STGAT using their 1V-1 model according to [13, 17]. We will also report the multimodal prediction results of our method, however, due to the limitation of space, these results will be shown in supplymentary file.

Method	ADE	FDE
w/o BiLSTM	0.51	1.04
ours	**0.48**	**0.99**

Table 2. Ablation study of BiLSTM for individual representation ($T_{pred} = 12$). Model in the first row uses LSTM as historical trajectory encoder instead of BiLSTM.

We will show our solid experiment results in Sec. 4.1, ablation study in Sec. 4.2, and qualitative analysis in Sec. 4.3.

4.1. Quantitive Analysis

Our method is evaluated on the popular ETH & UCY benchmark with ADE and FDE metrics for $T_{pred} = 12$. Experimental results is shown in Tab. 1. The results show that the performance of our model surpasses state-of-the-art methods on both ADE and FDE on most subsets. We reach an improvement of 11.1% and 10.8% in ADE and FDE in average respectively comparing with STGAT.

There is a special case that our method failed comparing with STGAT in UNIV dataset. The reason may be that there are a number of scenes in UNIV dataset where the number of pedestrians is huge (20 or more), while in other datasets this circumstances almost nonexist. When we apply a leave-one-out approach for training and evaluation on UNIV dataset, the RSBG generator will not be trained on huge groups but will be tested on these, which may lead to a performance degradation. Thus, this failure case may be caused by the unbalanced data distribution in leave-one-out test.

Note that the experiment results show that when human

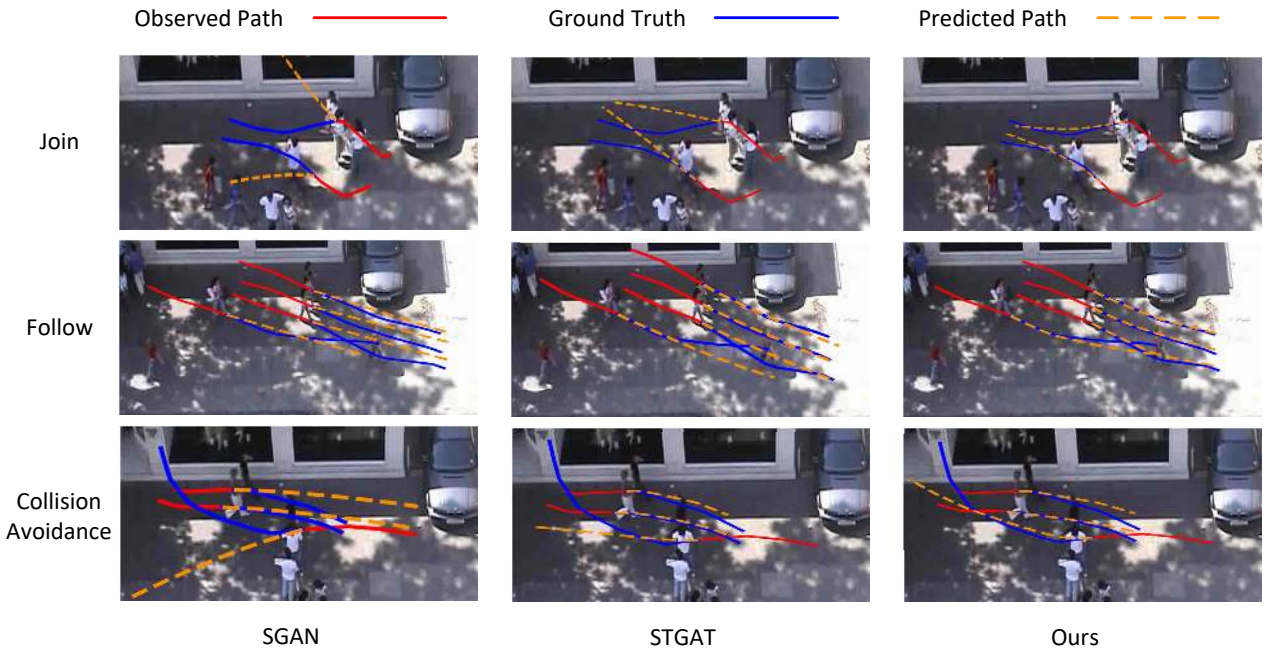

Figure 4. Comparisons between our model with STGAT(1V-1) and SGAN(1V-1) in three challenging social scenarios. We choose joining, following and collision avoidance here as three common social cases. For a better view, only key trajectories is presented.

context features are applied in our model, the performance will get worse in some subsets. This may also caused by the leave-one-out test since context feature changes a lot in different scenarios. Results in ETH dataset show that context features may be helpful for prediction in certain cases.

4.2. Ablation Study

BiLSTM encoder Comparing with most previous works [13, 17], we use BiLSTMs to encode historical trajectory of a single person rather than LSTMs, considering that later trajectories will influence the former ones as discussed in Sec. 3.3. To prove the effect of BiLSTM, we replace BiLSTM encoders by LSTM encoders in our model while other modules remain the same, and compare it with our full model. As shown in Tab. 2, BiLSTM encoders bring 5.9% in ADE and 4.8% in FDE improvement in average.

Exponential L2 Loss Because L2 Loss treats all timesteps in prediction phase as equivalent, it does not highlight enough on FDE while an accurate final position of a pedestrian is very important for trajectory prediction. Thus, we introduce Exponential L2 Loss to train the model. We represent four different settings of hyper parameter γ in Tab. 3 (∞ means using L2 Loss). By using a proper $\gamma = 20$, the average error rate is reduced by 4.0% and 4.8% for ADE and FDE in average respectively. However, if the loss overemphasize FDE by setting γ to small, it will bring an adverse effect according to the third row in Tab. 3.

Value	ADE	FDE
$\gamma = \infty$	0.50	1.04
$\gamma = 50$	0.49	1.01
$\gamma = 20$	**0.48**	**0.99**
$\gamma = 5$	0.52	1.06

Table 3. Ablation study for Exponential L2 Loss ($T_{pred} = 12$). We represent four various settings of hyper parameter γ here to show the influence of different degrees of emphasis on FDE. $\gamma = \infty$ means using L2 Loss.

4.3. Qualitative Analysis

Socially acceptable trajectory generation. One great challenge for human trajectory forecasting is to generate socially acceptable results as mentioned in [13]. Due to the diversity of social norms, we compare our methods with state-of-the-art approach STGAT and SGAN in three common social cases: joining, following and collision avoiding. Visualization results are shown in Fig. 4. We choose three challenging scenes that the slope of these trajectories changes frequently, which brings difficulties for prediction.

For joining case in row 1, our model successfully predict the fact that the man and the lady will join together after being separated by other pedestrians. SGAN do not capture this relation while prediction by STGAT gives a wrong joining direction and destination. The following scene in row 2 shows that our model have learned a common norm that

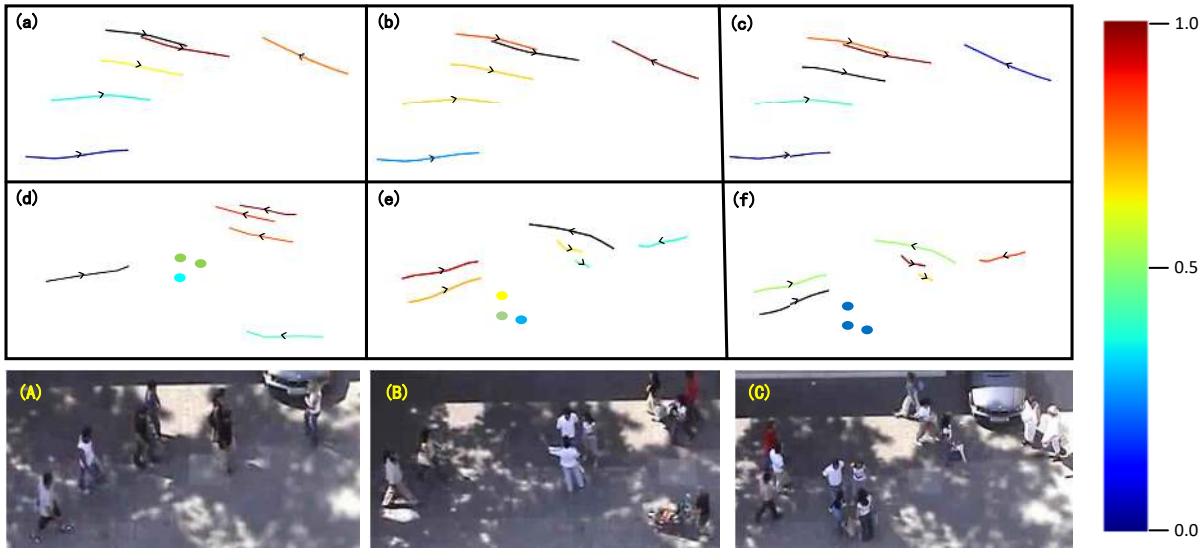

Figure 5. Figure (a)-(f) show relational social representation in RSBG. Different trajectories are marked by different colors and the direction is shown by arrows (Dots refer to pedestrians standing still). The range of color is from red to blue linearly, where red means strong relationship while blue means week relationship. The black trajectories are the target pedestrians. Figure (A)-(C) are real scenes corresponding to (a)-(c), (d), (e)-(f) respectively. Some pedestrians are not shown in RSBG because they are missing in the tracking files given by the dataset.

people are more inclined to following others if their starting point and destination are similar. Previous works do not exploit the latent social norm. Further, our model also gives a reasonable prediction in collision avoidance case in row 3. Although results from other methods avoid the conflict, predicted trajectories of the bottom agent point out that these models fail to predict his destination comparing with our method.

Social representation in RSBG. We visualize the social representation derived from RSBGs and analyze the latent group among these weights in Fig. 5. For a clear view, we show edge weights of key agents here.

Figure (a)-(c) show three relational social representation weights centered on three different person in the same scene. In this swarming and collision avoiding case, target person in (a) and (c) show a strong following tendency while target in (b) is more likely to avoid the collision, according to these visualized weights of edges in RSBG. This shows strong consistency with the behavior in our actual scenarios. Further, notice that the weights among these three targets are high, which infers that these three pedestrians are in a group.

Figure (d)-(f) show strong relationships between two distant pedestrians RSBG captured. In these three cases, the target agent gives more interest to those who he may have a conflict with rather than the pedestrians close to him. Particularly in case (f), RSBG figures out that there is an extremely high probability for the target person to collide with the approaching pedestrian even though he is the farthest one. These cases show that our method can successfully capture potential social relationships without influenced by the distance.

5. Conclusion

This paper studied human-human interactions among pedestrians for better trajectory prediction results. We proposed a novel structure called Recursive Social Behavior Graph, which is supervised by group-based annotations, to explore relationships unaffected by spatial distance. To encode social interaction features, we introduced GCNs which can adequately integrate information from nodes and edges in RSBG. Further, we used a plausible Exponential L2 Loss instead of common used L2 Loss to highlight the importance of FDE. We showed that by applying a group-based social interaction modeling, our model learns more latent social relations and performs better than distance-based methods.

6. Acknowledgement

This work is supported in part by the National Key R&D Program of China, No. 2017YFA0700800, National Natural Science Foundation of China under Grants 61772332 and Shanghai Qi Zhi Institute. We also acknowledge SJTU-SenseTime Joint Lab.

References

[1] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1. ACM, 2004.

[2] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016.

[3] Javad Amirian, Jean-Bernard Hayet, and Julien Pettré. Social ways: Learning multi-modal distributions of pedestrian trajectories with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[5] Niccoló Bisagno, Bo Zhang, and Nicola Conci. Group lstm: Group trajectory prediction in crowded scenarios. In *Proceedings of the European conference on computer vision (ECCV)*, pages 0–0, 2018.

[6] Chiho Choi and Behzad Dariush. Looking to relations for future trajectory forecast. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[7] Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. End-to-end continuous speech recognition using attention-based recurrent nn: First results. *arXiv preprint arXiv:1412.1602*, 2014.

[8] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[9] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. Pairwise body-part attention for recognizing human-object interactions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 51–67, 2018.

[10] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2334–2343, 2017.

[11] Hao-Shu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[12] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. Graph convolutional tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4649–4659, 2019.

[13] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018.

[14] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995.

[15] Mikael Henaff, Joan Bruna, and Yann LeCun. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*, 2015.

[16] Tsubasa Hirakawa, Takayoshi Yamashita, Toru Tamaki, and Hironobu Fujiyoshi. Survey on vision-based path prediction. In *International Conference on Distributed, Ambient, and Pervasive Interactions*, pages 48–64. Springer, 2018.

[17] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[18] Boris Ivanovic and Marco Pavone. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2375–2384, 2019.

[19] Chanho Kim, Fuxin Li, and James M. Rehg. Multi-object tracking with neural gating using bilinear lstm. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[20] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In *European Conference on Computer Vision*, pages 201–214. Springer, 2012.

[21] Laura Leal-Taixé, Michele Fenzi, Alina Kuznetsova, Bodo Rosenhahn, and Silvio Savarese. Learning an image-based motion context for multiple people tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3542–3549, 2014.

[22] Yuke Li. Which way are you going? imitative decision learning for path forecasting in dynamic scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[23] Yong-Lu Li, Liang Xu, Xijie Huang, Xinpeng Liu, Ze Ma, Mingyang Chen, Shiyi Wang, Hao-Shu Fang, and Cewu Lu. Hake: Human activity knowledge engine. *arXiv preprint arXiv:1904.06539*, 2019.

[24] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *CVPR*, 2019.

[25] Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander G Hauptmann, and Li Fei-Fei. Peeking into the future: Predicting future person activities and locations in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5725–5734, 2019.

[26] Wei-Chiu Ma, De-An Huang, Namhoon Lee, and Kris M. Kitani. Forecasting interactive dynamics of pedestrians with fictitious play. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[27] Yuexin Ma, Xinge Zhu, Sibo Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6120–6127, 2019.

[28] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model.

In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 935–942. IEEE, 2009.

[29] Bo Pang, Kaiwen Zha, Hanwen Cao, Chen Shi, and Cewu Lu. Deep rnn framework for visual sequential applications. In *CVPR*, 2019.

[30] Bo Pang, Kaiwen Zha, Yifan Zhang, and Cewu Lu. Further understanding videos through adverbs: A new video task. In *AAAI*, 2020.

[31] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th International Conference on Computer Vision*, pages 261–268. IEEE, 2009.

[32] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European conference on computer vision*, pages 549–565. Springer, 2016.

[33] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 300–311, 2017.

[34] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. 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*, pages 1349–1358, 2019.

[35] Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–7. IEEE, 2018.

[36] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[37] Runzhong Wang, Junchi Yan, and Xiaokang Yang. Learning combinatorial embedding networks for deep graph matching. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[38] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *arXiv preprint arXiv:1901.00596*, 2019.

[39] Dan Xie, Sinisa Todorovic, and Song-Chun Zhu. Inferring "dark matter" and "dark energy" from videos. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[40] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S. Kankanhalli. Learning to detect human-object interactions with knowledge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[41] Hao Xue, Du Q Huynh, and Mark Reynolds. Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1186–1194. IEEE, 2018.

[42] Kota Yamaguchi, Alexander C Berg, Luis E Ortiz, and Tamara L Berg. Who are you with and where are you going? In *CVPR 2011*, pages 1345–1352. IEEE, 2011.

[43] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[44] Bolei Zhou, Xiaogang Wang, and Xiaoou Tang. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2871–2878. IEEE, 2012.