

Learn to Quantify Social Interaction with Constraints for Pedestrian Walking

Xiaodan Shi

Abstract—Long-term human path forecasting in crowds is critical for autonomous moving platforms (like autonomous driving cars and social robots) to avoid collision and make high-quality planning. Although the current research take into account social interactions for prediction, they don't reveal the exact kinds of social interactions happened among people and how the social interactions affect the decision-making process of pedestrians, which further limits its robustness. Social interactions in pedestrian walking are intuitively massive and hard to label and quantify. In this paper, we explore creatively to quantify and interpret how pedestrians interact with others by proposing *Learn to Cluster*. Our clustering social interactions is probabilistic latent variable generative, learning directly from sequential trajectory observations, scalable to arbitrary number of pedestrians. Learn to cluster is label-free and can be naturally integrated into the training process of the prediction model. The latent variables will then serve as 'labels' to categorize social interactions. Extensive experiments over several trajectory prediction benchmarks demonstrate that our method is able to learn the patterns of social interactions and effectively integrate the patterns to pedestrian trajectory prediction.

Keywords: social interactions, trajectory prediction, learn to cluster, interpretability, autonomous driving

I. INTRODUCTION

Learning social etiquette lies at the heart of understanding human trajectories in crowded scenes. For example, the pedestrians or bicyclists can quickly integrate all the surrounding pedestrians, bicyclists or cars and yield the right-of-the-way. Pedestrians naturally have learnt the ways how to interact with physical space and other moving objects by avoiding collisions and respecting the social distance. Learning those principles are not only important for understanding pedestrians behaviors but also helps autonomous vehicles and robots to navigate safely in very complex scenes with extraordinary proficiency.

Understanding how pedestrian socially interact with others in crowds is at the core of trajectory prediction problem. Trajectory prediction is a problem that predicts future positions of pedestrians or vehicles given past trajectory observations, has raised much attentions in recent years due to its potential applications on many transportation applications. The models of trajectory prediction are usually agent-based, tackling the social interactions to make more accurate predictions by integrating the hidden moving states of surrounding objects through a graph structure.

Social interactions among crowds are complicated, heterogeneous and hard to interpret. Despite a large body of works

Xiaodan Shi is with Department of Computer and Systems Sciences, Stockholm University xiaodan.shi@dsv.su.se
Corresponding Author: Xiaodan Shi

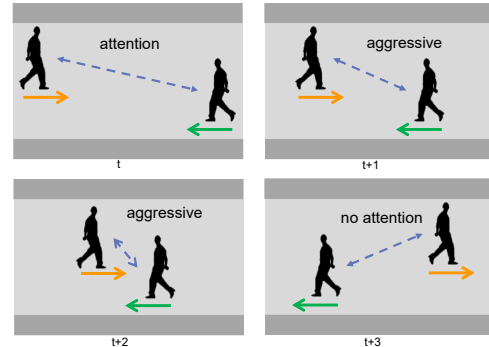


Fig. 1. From a common-sense perspective, social interactions can be broadly categorized into aggressive, mild, and no interaction. The scenarios illustrated represent relatively simple cases, in which the presence or absence of social interaction can often be readily identified. However, in more complex scenarios, it becomes much more difficult to determine whether social interactions occur.

on modeling social interactions for trajectory prediction, the existing research usually visually analyse the social behaviors in afterwards experiments as avoid collisions, walking together and merging into a group etc. However, no exclusive information about the types of social interactions are truly considered in networks. Trajectory prediction models usually blindly pool all kinds of social interactions without considering their distinction, which means the networks modeling social interactions are not interpretable. The types of social interactions and how they affect the decision-making process of pedestrians aren't revealed yet through the current research.

In this paper, we shed light on the interpretability of social interactions for trajectory prediction, creatively revealing how pedestrians exactly socially interact with others and their decision-making process for navigation through crowds. To achieve this, a unifying framework is proposed to describe the possible types of social interactions. Our framework modeling social interactions is probabilistic latent variable generative, learning directly from sequential trajectory observations, scalable to arbitrary number of pedestrians. Our model takes on the point-of-view of any agent and aggregates the relative motions between agent and its neighbor pedestrians through a common-used graph structure. Latent variables further automatically learns to describe the possible socially meaningful modes of social interactions in a probability way without explicit labeling. In fact, our modeling social interactions clusters all the social interactions through the discrete latent variables, where social behaviors in the same cluster should be similar and in the different cluster should

be distinctive. To capture this, we also propose a social loss function to regulate the inter- and cross-cluster "distances" to maximize the cross-cluster similarities and to minimize the inter-cluster similarities.

Besides, we also propose a simple yet efficient method for integrating the learned modes of social interactions to prediction model. In the prediction model, we adopt Mixture Density Network (MDN) to generate multiple future positions and compute the loss over all components of the mixture model for capturing the uncertainty of prediction. We test the model using classic trajectory prediction benchmarks. To address the challenge of interpreting the learned clusters, we propose an intuitive strategy by designing several common-sense interaction scenarios to indirectly explain the meaning of these modes. The experiments show promising and comparable results.

II. RELATED WORKS

A. social Interactions for Trajectory Prediction

Social LSTM introducing Social Pooling to learn a global feature of all nearby neighbors around an agent which is meant to represent common sense rules and social conventions, is a tipping point for data-driven long-term trajectory prediction. Many research follow the way of Social LSTM [1] but with improvements. Attention mechanism is introduced to learn neighbors' weights on agent [2], [3], [4]. Fernando et al. extended the classic model to incorporate both soft attention as well hard attention where the former is for handling longer trajectories and the latter is used for modeling interacting people [5]. Instead of directly modeling hidden states of neighbors' motion, some research pool relative motion between agent and neighbors to model interactions. SR-LSTM proposed a state refinement module for LSTM, which extracting social effects of neighbors by embedding and aggregating the relative spatial location between agent and neighbors [6]. Graph representation, specifically spatio-temporal graph (ST-graph) is well applied to illustrate human motion and their interactions [7], [8], [9], [10]. ST-graph provide a more direct and natural way to model interactions for trajectory prediction. Structure-RNN [11] combining high-level spatio-temporal graphs with sequence modeling success of RNN made significant improvements on problem of human motion modeling. Some research follow this direction. Social-BiGAT introduced a flexible graph attention network to model social interactions between pedestrians in a scene. It assumes all people in a scene interacting instead of setting a local neighborhood [12]. Social-STGCNN utilized spatio-temporal graph representation and proposed a weighted adjacency matrix to measure the influence between pedestrians [8]. Recently, Transformer is also used to model the motion and social interactions for trajectory prediction [13], [14], [15]. Li etc. utilized self-attention mechanism to integrate social interactions by using queries Q to represent the agent actor, keys K and values V to represent neighbor agents [13]. Although most of the current research claim they consider social interactions for future prediction, it is hard to say what kind of social interactions

going on among pedestrians are really encoded. Thus in the paper, we investigate to explain the social netiquettes among pedestrians and to encode the explainable social interactions for prediction problem.

B. Multi-modality of Trajectory Prediction

Human motions under crowded scenarios imply a multiplicity of modes. To capture the uncertainty of future path, some research apply generative adversarial network (GAN) or variable autoencoder (VAE) to generate multiple possible paths [16], [2], [17], [18], [19]. Gupta A. et al. proposed Social GAN which contains RNN based encoder-decoder generator and RNN-based decoder discriminator [16]. Social GAN integrates all the interactions involved in the scenarios and encourages the generative network to spread its distribution and cover the space of possible paths by introducing a variety loss. Sadeghian A. et al proposed Sophie, an attentive GAN to jointly model static human-space, and dynamic human-human interactions by blending a social attention mechanism with a physical attention that helps the model to learn where to look in a large scene and to extract the most salient parts of the image relevant to the path [2]. Some research apply Mixture Density Network (MDN) to map the distribution of future trajectories [7], [20], [21], [22]. The article [21], based on MDN, proposed a two stage strategy that first predicted several samples of future with Winner-Takes-All loss and then iteratively grouped the samples to multiple modes. There are also goal-based multi-trajectory prediction [23], [24], [25], [26], [27]. Those models predict multiple futures based on hypothesis of goals. One kind of goal-based prediction models the trajectories based on the semantic destinations, such as turning right/left, going straight [23], [28]. Another kind firstly forecasts multiple positional designations and then estimates futures matching the goal hypothesis [29]. We also model the multi-modality of trajectory and forecast multiple plausible futures by using MDN. But worth noting that it is not our key contribution and we mainly focus on modeling explainable social interactions. We predict multiple futures mainly for: (1) to better compare our method with other baselines; (2) to demonstrate the proposed explainable social interactions able to apply to forecast multi-modal futures.

III. METHODOLOGY

A. Problem Formulation

We assume that each scenario has been preprocessed to get 2D spatial coordinates $(x_i^t, y_i^t) \in \mathbf{R}$ and 2D walking speed $(u_i^t, v_i^t) \in \mathbf{R}$ of all people at all time instances. There are N agents in a scenario. The observation of agent i is past trajectories represented as: $X_i^{1:\tau} = \{(x_i^t, y_i^t, u_i^t, v_i^t) | t = 1, 2, \dots, \tau\}$ while the future trajectories is $Y_i^{\tau:T} = \{(x_i^t, y_i^t) | t = \tau + 1, \dots, T\}$.

Our goal is to learn the posterior distribution $p(Y_i^{\tau:T} | X_i^{1:\tau}, X_{1:N \setminus i}^{1:\tau})$. To generate the distribution of future trajectories, we jointly model multiple ego-trajectories and their interactions with f . Therefore, the distribution is

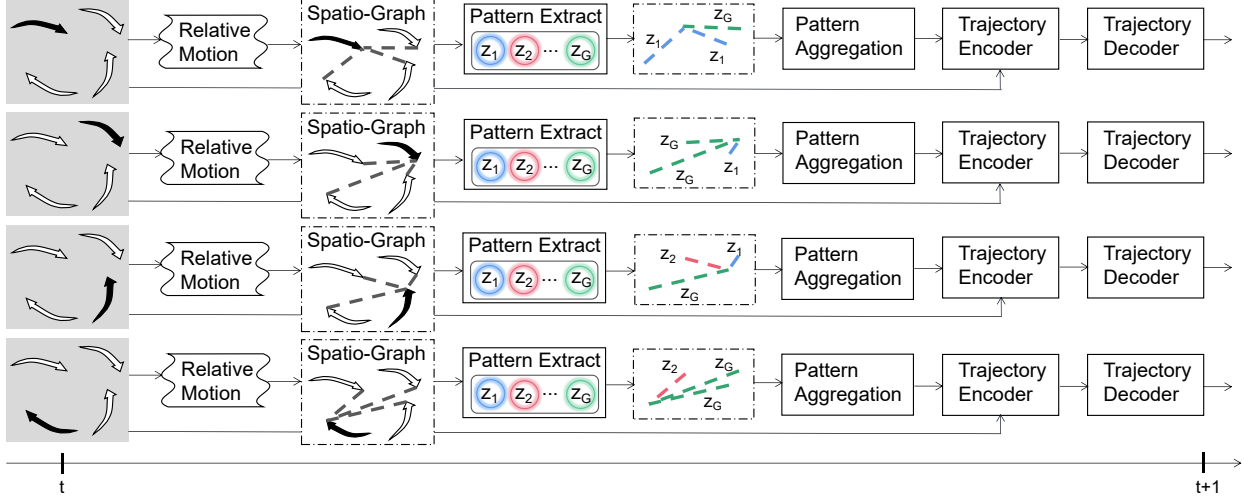


Fig. 2. Model Architecture. We first employ a spatio-temporal graph [8] to jointly represent pedestrian interactions and walking trajectories. Within the prediction network, latent variables are learned from the features of social interactions to capture their clustered representations. The social interaction features, together with the corresponding cluster information, are then integrated into the prediction process through a pattern aggregation module.

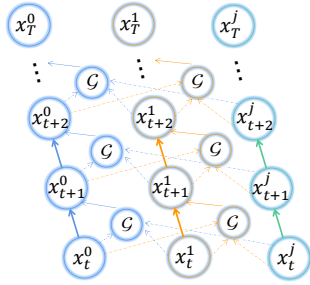


Fig. 3. Calculation of mode. We calculate the mode of social interactions at every time instance.

denoted as:

$$p(Y_i^{\tau:T} | X_i^{1:\tau}, X_{1:N \setminus i}^{1:\tau}) = f(X_i^{1:\tau}, X_{1:N \setminus i}^{1:\tau}; w^*), \quad (1)$$

where w^* are the parameters of the model we aim to learn. We denote the predicted future paths as $\hat{Y}^{\tau:T}$ whose distributions are learned from our model.

B. Social Interactions

As shown in Figure 2, all the pedestrians at any time instance t are characterized with location (x^t, y^t) and offset (u^t, v^t) . Offsets are common-used states describing trajectories, which can stabilize the training process and improve the model performance. We assume the agent pedestrian is indexed as i and its social graph is constructed by formulating relative motions between it and all the neighbors in a neighborhood setting. We use \mathcal{E}_j^t to denote the spatial edge between i and neighbor j as follows:

$$\mathcal{E}_{i,j}^t = [(x_i^t, y_i^t) - (x_j^t, y_j^t), (u_i^t, v_i^t) - (u_j^t, v_j^t)] \quad (2)$$

We assume there are M neighbors of the agent at time instance t . $\mathcal{E}_i^t = \{\mathcal{E}_{i,j}^t | j = 1, 2, \dots, M\}$, has the size of $(M, 4)$. M is dynamic, free to be any number as long as the neighbor pedestrians are in the neighborhood of the agent. $\mathcal{E}_{i,j}^t$ is then embedded with a fully connected layer with ReLU activation and featured as $\hat{\mathcal{E}}_{i,j}^t$.

To interpret and extract patterns of social interactions, we introduce a set of stochastic latent variables $\mathcal{G} \sim \text{Multinoulli}(G)$ that is directly embedded in the prediction network and conditioned on the latent features of social interactions and will learn to indicate the similarity and discrimination between massive dynamic interactions. For each pairwise agents at time instance t , $\mathcal{G}_{i,j}$ is conditioned on $\hat{\mathcal{E}}_{i,j}^t$ and can take on G discrete values to define the social interaction patterns of the spatial edge (Figure 3). Those discrete variables are finite and latently label each social interaction directly from the trajectory data. We understand social interactions with the same latent labels lie within the same cluster, sharing the similar social behaviours, owning the same interaction pattern. Likewise, social interactions with different variables have different social behaviours. The social interactions patterns/clusters described by variable \mathcal{G} don't match the explicit meanings initially. The patterns don't mean the exact social behaviours such as merging into a group, walking parallel etc., but tend to indicate the styles of social behaviours such as aggressive: agents pay attention to those neighbors, try to avoid collision with them and adjust path mainly based on their interactions, mild: agents pay attentions to those neighbors, keep safe distance with them and adjust path partly based on their interactions, no attention: pay no attention to those neighbors, etc. As we mentioned before, we further construct a tensor $\hat{\mathcal{E}}_i^t$ by embedding each element of \mathcal{E}_i^t to more succinctly

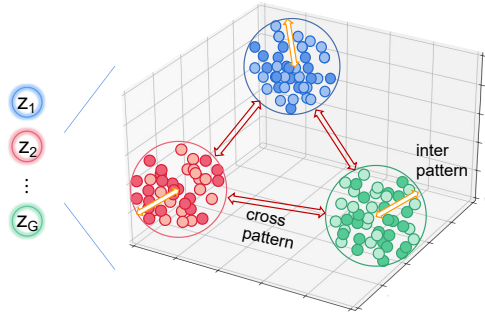


Fig. 4. Regulations on social patterns. The features of same clusters contract and features of different clusters contrast.

get the embedding features of our latent patterns by using a trick similar to how Transformer model get multi-head attentions. $\hat{\mathcal{E}}_i^t$ has original size of $(M, R * G)$ where $R * G$ is the embedding size of pairwise relative motion $\hat{\mathcal{E}}_{i,j}^t$, is then reshaped as (M, G, R) . We use a softmax function to $\hat{\mathcal{E}}_i^t$ to get discrete probability distribution of social interaction patterns as follows:

$$p(\mathcal{G}_i^t | X_{1:M}^t) = \text{Softmax}(\hat{\mathcal{E}}_i^t) \quad (3)$$

where, $p(\mathcal{G}_i^t | X_{1:M}^t)$ is the probability distribution over G interaction patterns and reflects the possibilities of patterns each social interaction belong to. To get the exact social pattern, we need to sample from the discrete probabilities. However, the sampling process in the middle of the network lead to the problem of backpropagation interruption due to its non-differentiability. To address it, we further utilize Gumbel-Max, the reparameterization trick to sample from $p(\mathcal{G}_i^t | X_{1:M}^t)$.

$$\mathcal{G}_i^t = \text{argmax}(\log(p(\mathcal{G}_i^t | X_{1:M}^t)) + g) \quad (4)$$

where, $g = -\log(-\log(u))$, $u \sim \text{uniform}(0, 1)$ is the Gumbel noise. The function $\text{argmax}(\cdot)$ is not differential, we further use $\text{softmax}(\cdot)$ to approximate it. \mathcal{G}_i^t has the shape (M, G) of which each row is an one-hot vector where element 1 describe the social interaction pattern.

C. Regulations on Social Patterns

Real-world trajectory-related social behaviors vary on-the-fly and are hard to interpret. Our hypothesis social patterns can be learned from trajectory prediction directly and then help prediction in turn. At the core of our approach is discriminating the social features in a representation space as shown in Figure 4. Features of similar social behaviors should be clustered together and far away from other social behaviors. Shortly speaking, the features of same clusters contract and features of different clusters contrast. Inspired by Maximum Coding Rate Reduction [30], we utilize coding rate to describe the compactness of the features of the clusters. More specifically, the entire space of features from all clusters should span a space of the largest volume and the coding rate of whole feature set should be as large as

possible while the coding rate of the features of the same clusters should be as small as possible.

By expanding the size of $\hat{\mathcal{E}}_i^t$ to (M, G, R) and multiplying it with \mathcal{G}_i^t , social features are gathered for each patterns. The process is repeated for each agent pedestrian. After tensor rearrangement, we obtain $Z^t = \{z_m^t | m = 1, 2, \dots, G\}$ where z_m represent all the social features belong to pattern m . The number of features stored in z_m is dynamic during training and testing. According to [31], the average coding rate of features of all clusters at any time instance t can be calculated as:

$$R_c(Z^t, \epsilon | \Pi) = \sum_{m=1}^G \frac{\text{tr}(\Pi^m)}{2M} \log \det \left(\mathbf{I} + \frac{d}{\text{tr}(\Pi^m)\epsilon^2} Z^t \Pi^m Z^{t*} \right) \quad (5)$$

Likewise, the coding rate of the entire feature space is calculated as:

$$R(Z^t, \epsilon) = \frac{1}{2} \log \det \left(\mathbf{I} + \frac{d}{M\epsilon^2} Z^t Z^{t*} \right) \quad (6)$$

where, $\Pi = \{\Pi^m \in R^{(M,M)}\}_{m=1}^G$ is a set of diagonal matrices whose diagonal entries represent the cluster to which the social interaction between agent and pedestrian M belongs. $\text{tr}(\cdot)$ is the trace of a square matrix, ϵ is the distortion. The loss item based on $R_c(Z^t, \epsilon | \Pi)$ and $R(Z^t, \epsilon)$ is:

$$\mathcal{L}_{\text{social}} = - \sum_{t=\tau}^{T-1} \lambda (R(Z^t, \epsilon) - R_c(Z^t, \epsilon | \Pi)) \quad (7)$$

$\mathcal{L}_{\text{social}}$ allows to learn larger $R(Z^t, \epsilon)$ meaning more distinct cross patterns while smaller $R_c(Z^t, \epsilon | \Pi)$ meaning a more focused inner pattern, which is preferred in our method.

D. Pattern Aggregation

A simple yet efficient aggregation strategy is introduced by combining social features and social patterns for path forecasting. Firstly, we encode the social patterns formatted as one-hot vectors. Practically, a fully connected layer is used for \mathcal{G}_i^t of shape (M, G) to get $\hat{\mathcal{G}}_i^t$ of shape (M, R) . Follow the symbol usage of $\hat{\mathcal{E}}_i^t$, we use $\hat{\mathcal{G}}_i^t$ to denote social features of agent i . By a multiplication then a sum up operation between $\hat{\mathcal{E}}_i^t$ and $\hat{\mathcal{G}}_i^t$ over each corresponding pattern, an aggregated feature η_i^t specified for agent i is obtained.

E. Path Forecasting

As mentioned before, the agent i at time instance t is characterized with location (x_i^t, y_i^t) and offset (u_i^t, v_i^t) . Following most of the current research, we use offset for trajectory prediction, which not only stabilize the whole training process but also improve the prediction performance. A fully connected layer with ReLU non-linearity is applied to (u_i^t, v_i^t) and get features f_i^t . f_i^t and η_i^t is added then for predicting the offset state of next time step.

$$h_i^t = \phi(h_i^{t-1}, f_i^t + \eta_i^t; w_h^*) \quad (8)$$

where, $\phi(\cdot)$ is the trajectory encoder. Here, we use LSTM and its weights w_h^* are shared between all people in a scenario.

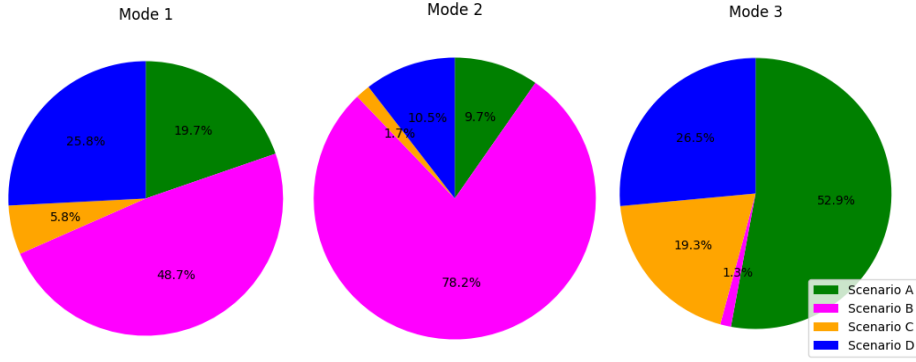


Fig. 5. Interpretation on Mode. We explain modes by analyzing them from several typical interaction scenarios that are designed based on common sense.

Trajectory decoder also share the same weight with encoder while encoder is for encoding trajectory observations and decoder is for futural prediction. To capture the multi-modality of future paths, we utilize mixture density network(MDN) that combines a multilayer perception with GMMs. The next location of agent conditioned on hidden states of LSTM h_i^t are denoted as:

$$p(\hat{Y}_i^{t+1}|h_i^t) = \sum_{g=1}^B \alpha_g^t p(\hat{Y}_i^{t+1}|\mu_g^t, \sigma_g^t) \quad (9)$$

where B is the number of Gaussian models of MDN, α_g^t is the prior of g th kernel, $p(\hat{Y}_i^{t+1}|\mu_g^t, \sigma_g^t)$ is the probability density functions (PDFs) given by g th component of GMMs which is a bivariate Gaussian model parametrized by the mean $\mu_g^t = (\mu_x, \mu_y)^t$, standard deviation $\sigma_g^t = (\sigma_x, \sigma_y)^t$ and correlation coefficient ρ_g^t . We set ρ_g^t as constant and learn μ_g^t , σ_g^t and α_g^t through our network.

$$\begin{aligned} \alpha_g^t &= \frac{\exp(a_g^t)}{\sum_{k=1}^M \exp(a_k^t)} \\ \mu_g^t &= u_g^t \\ \sigma_g^t &= \exp(z_g^t) \end{aligned} \quad (10)$$

where $\{a_g^t|g = 1, \dots, B\}$, $\{u_g^t|g = 1, \dots, B\}$ and $\{z_g^t|g = 1, \dots, B\}$ is obtained by applying fully connected layers $\phi_\alpha(\cdot)$, $\phi_\mu(\cdot)$ and $\phi_\sigma(\cdot)$ to h_i^t respectively.

F. Loss Function

The loss function is designed to compute negative log-likelihood of future trajectories over all components of a mixture model (Eq. (10)) plus social loss.

$$\begin{aligned} \mathcal{L}_{mdn} &= -\sum_{t=\tau}^{T-1} \log(\sum_{g=1}^M \alpha_g^t p(\hat{Y}^{t+1}|\mu_g^t, \sigma_g^t)) \\ \mathcal{L}_{all} &= \mathcal{L}_{mdn} + \beta \cdot \mathcal{L}_{social} \end{aligned} \quad (11)$$

IV. EXPERIMENTS

In this section, we try to answer 3 questions: (1) what do modes mean, (2) when do modes shift, (3) will the social interaction clustering improve the prediction accuracy. Moreover, the proposed model is evaluated on two publicly available datasets: UCY [32] and ETH [33]. The two datasets contain 5 sets, which are UCY-zara01, UCY-zara02, UCY-univ, ETH-hotel, ETH-eth in 4 crowded scenarios with totally 1536 trajectories. We firstly preprocess those two datasets by

resampling them as 2.5fps and transforming the coordinates of people to world coordinates in meters.

Implementation Details. The experiments are implemented using Pytorch under Ubuntu 16.04 LTS with a GTX 4090 GPU. The size of hidden states of LSTM is set to 128. The embedding layers are composed of a fully connected layer with size 128. The batch size is set to 8 and all the methods are trained for 200 epochs. The optimizer RMSprop is used to train the proposed model with learning rate 0.001. We clip the gradients of LSTM with a maximum threshold of 10 to stabilize the training process. The model outputs GMMs with five components. β is set as 0.1. Specifically, we set G as 3 which is also compliant with the other research [34], [27], [35] which state that while in the ideal case of having infinite rationality, humans are expected to make better decisions through more complex cognition, the limited time, knowledge and computational power of humans enhances the desirability of simple cognition for more robust performance to cope with the uncertainty of the world. In the context of crowd walking, where rapid decision-making is required, we argue that social interactions tend to fall into only a few categories.

Evaluation Approach. The proposed model is trained and tested on the two datasets with leave-one-out approach: trained on four sets and tested on the remaining set. We observe the trajectories for 8 timesteps (3.2 sec) and show prediction results for 12 timesteps (4.8 sec). To evaluate the performance, we compare our method with other state-of-the-art models on two generally used metrics.

1. Average displacement error (ADE): average L2 distance over all prediction results and ground truth. ADE measures average error of the predicted trajectory sequence.

2. Final displacement error (FDE): distance between prediction result and ground truth at final timestep. FDE measures the error "destination" of the prediction.

A. What do modes mean?

Since the clustering of social interactions is performed without labels, interpreting the meaning of each mode is inherently challenging. To provide an intuitive explanation of the clustered categories, we define several typical interaction scenarios based on common sense. **Scenario A** describes

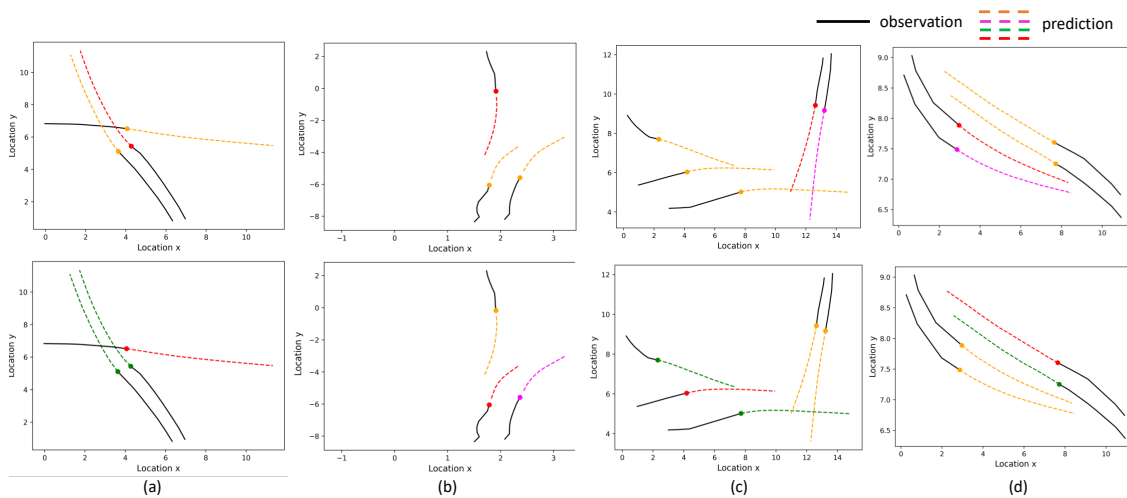


Fig. 6. Visualization on modes. Mode 1-green, Mode 2-Magenta, Mode 3-Orange. The red dot and red lines represent the agent pedestrian while the other neighbor pedestrians' color show their social interaction mode with the agent which is how agent socially interacted with his/her neighbors.

two pedestrians approaching each other from opposite directions at a close distance, a situation that is more likely to involve intense and aggressive interaction. **Scenario B** corresponds to two pedestrians walking in parallel in the same direction at a close distance, which typically reflects following behavior and can be regarded as mild interaction or even non-interaction. **Scenario C** represents two pedestrians approaching from opposite directions but starting at a relatively large distance, which tends to result in all kinds of social interactions which are mild interaction, no interaction or aggressive social interaction. **Scenario D** describes two pedestrians walking back-to-back in opposite directions with a large separation, a situation that usually indicates no interaction. We performed a statistical analysis of social interactions in the dataset. The results are presented in the Figure 5. It can be observed that **Mode 3** is predominantly distributed in Scenario A, indicating that social interactions belonging to this mode are more likely to involve **aggressive interactions**. In contrast, **Mode 2** is more concentrated in Scenario B, suggesting that it tends to exhibit **mild forms of interaction**. It is worth noting that **Mode 1** is relatively evenly distributed across all four scenarios, implying that this mode represents a more common form of interaction that is less influenced by pedestrians' relative positions or walking directions. Therefore, we recommend interpreting mode 1 as corresponding to **non-interaction** or very slight social interactions.

To further gain insight into the meaning of each mode, i.e., the interpretation of the clusters, we conducted a visualization across all datasets, with several representative examples shown in the Figure 6. We observe that social interactions within a group are not always symmetric. For instance, in Figure (a1) (the upper subfigure of panel a), the agent exhibits aggressive interactions with its two neighbors, while in Figure (a2) the agent is already moving away from its neighbors and therefore pays less attention to them. In Figure

(b2), the agent primarily attends to the pedestrian on its left, but at the same time interacts with the neighbor on the right as it intends to make a right turn. The scenarios in Figures (c) and (d) involve a larger number of pedestrians, yet the interaction behaviors remain consistent with those observed in other cases. This further demonstrates the rationality of the learned modes by our proposed method.

B. When do modes shift?

We further analyze the short time periods when a mode transition occurs, aiming to understand which relational factors between the agent pedestrian and its neighbors change, thereby providing insights into the elements that drive the transition. Due to space limitations, we only present the cases of Mode 2 \rightarrow Mode 3 and Mode 3 \rightarrow Mode 1 in Figure 7. We design four indicators for this analysis: (1) relative speed magnitude, (2) walking direction (i.e., same, opposite, crossing), (3) relative distance, and (4) the relative position of neighbors with respect to the agent (i.e., front, back, side).

Our analysis reveals that when the social interaction becomes more intense (e.g., Mode 2 \rightarrow Mode 3), the most significant change lies in the relative positional relationship, as more neighbors appear in front of the agent. At the same time, the relative speed magnitude increases and the relative distance decreases, while the relative walking direction does not change substantially. In contrast, when the social interaction gradually becomes milder (e.g., e \rightarrow f in Figure 7), more neighbors move to the rear of the agent, and the walking direction shifts more often from opposite to same or crossing. Correspondingly, the relative distance slightly increases, while the relative speed magnitude slightly decreases. Based on the above analysis, we may argue that the relative positional relationships between pedestrians are one of the main factors driving mode transitions.

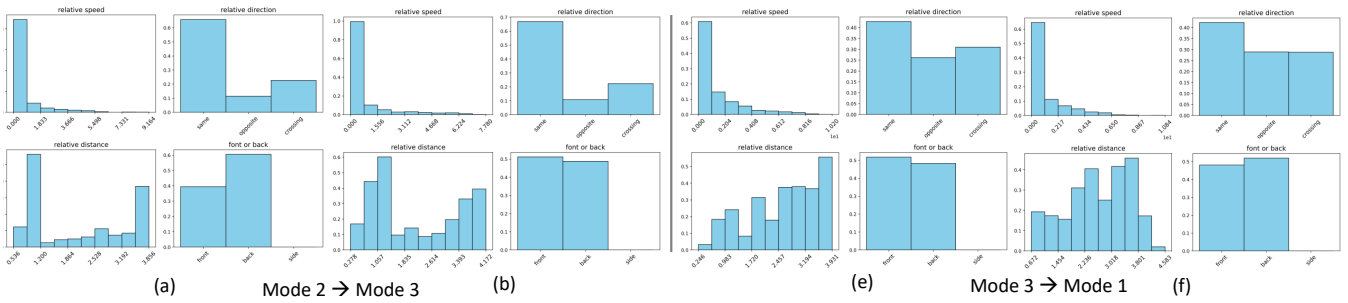


Fig. 7. Statistics on mode change. We analyze the change on relative speed(upper left), walk direction(upper right), distance(lower left) and position relation(lower right) respectively when mode change from 2 to 3 and from 3 to 1.

TABLE I

QUANTITATIVE RESULTS OF BASELINES VS. OUR METHOD ACROSS DATASETS FOR PREDICTING 12 FUTURE TIMESTEPS(4.8 SEC) GIVEN 8 TIMESTEPS OBSERVATION(3.2 SEC).

Method	Note	Evaluation (ADE(m)/FDE(m))					
		ETH-eth	ETH-hotel	UCY-univ	UCY-zara01	UCY-zara02	AVG
Linear	kalman filter	1.65/2.34	0.99/1.70	0.86/1.51	0.83/1.44	0.54/0.96	0.97/1.69
LSTM	offset as input	0.71/1.40	1.15/2.09	0.72/1.49	0.48/0.98	0.38/0.77	0.69/1.35
Social LSTM [1]	social pooling	1.09/2.35	0.79/1.76	0.67/1.40	0.47/1.00	0.56/1.17	0.72/1.54
Sophie[2]	20 samples	0.70/1.43	0.76/1.67	0.54/1.24	0.30/0.63	0.38/0.78	0.54/1.15
Social GAN[16]	20 samples	0.72/1.29	0.48/1.01	0.56/1.18	0.34/0.69	0.31/0.65	0.48/0.96
Social BiGAT[36]	20 samples	0.69/1.29	0.49/1.01	0.55/1.32	0.30/0.62	0.36/0.75	0.48/1.00
Social STGCNN[8]	20 samples	0.64/1.11	0.49/0.85	0.44/0.79	0.34/0.53	0.30/0.48	0.44/0.75
Model-V1	20 samples	0.33/0.68	0.16/0.36	0.60/1.22	0.30/0.66	0.27/0.45	0.33/0.67

C. Mode Accuracy

To further illustrate the significance of learning clustered social interactions, we also evaluate their impact on trajectory prediction accuracy. Table I presents a comparison of the proposed method against baseline approaches in terms of prediction accuracy. Linear methods underperform because they cannot capture social context or the multi-modality of human motion. Social LSTM performs on par with a vanilla LSTM, while offset-based LSTM inputs stabilize training and improve accuracy. Methods such as Sophie, Social GAN, and Social BiGAT, which model long-term uncertainty, achieve superior results. “20 samples” denotes selecting the best outcome from 20 samples drawn from the predicted distribution. Interpretability inevitably reduces the accuracy of prediction models. However, surprisingly overall, our model outperforms the baselines by a large margin on the ETH datasets, demonstrating its strong capability to capture complex interaction dynamics in crowded and unconstrained environments. On the UCY datasets, our method achieves performance comparable to other state-of-the-art baselines. These results highlight the robustness and generalizability of the proposed model across different benchmark datasets.

D. Ablation study

To explain how our model works, we also represent results (Table 2) of various versions of our models in an ablative setting by Model-V1 and Model-V2: the whole version of our model with different sampling times, Model-V3: with the entire model, mdn loss without social loss, i.e. without regulations on modes, Model-V4: with the entire loss and uses

TABLE II

ABLATION STUDY ON OUR MODEL. M INDICATE IF WE LEARN MODE FOR SOCIAL INTERACTIONS OR NOT, K MEANS SAMPLING TIMES, LOSS PRESENT IF WE USE SOCIAL LOSS OR NOT.

Variant ID	Components			Evaluation (ADE(m)/FDE(m))					
	M	K	Loss	ETH-eth	ETH-hotel	UCY-univ	UCY-zara01	UCY-zara02	AVG
Model-V1	✓	20	\mathcal{L}_{all}	0.33/0.68	0.16/0.36	0.60/1.22	0.30/0.66	0.27/0.45	0.33/0.67
Model-V2	✓	1	\mathcal{L}_{all}	0.46/1.12	0.36/1.00	0.70/1.51	0.44/1.09	0.32/0.76	0.46/1.10
Model-V3	✓	1	\mathcal{L}_{mdn}	0.47/0.90	0.26/0.46	0.68/1.34	0.37/0.68	0.29/0.51	0.41/0.78
Model-V4	✓	1	\mathcal{L}_{mdn}	0.77/1.68	0.47/0.92	0.74/1.51	0.58/1.23	0.36/0.76	0.58/1.22
Model-V5	✓	1	\mathcal{L}_2	0.60/1.21	0.39/0.95	0.69/1.50	0.47/1.14	0.40/0.80	0.51/1.12

attention mechanism to generate social states without mode extraction, Model-V5: with Mean Squared Error (MSE) loss, with mode extraction and applies softmax over each mode before integration.

By comparing Model-V2 and Model-V3, we observe that introducing constraints into the clustering process is beneficial, as it significantly improves prediction accuracy. Moreover, such constraints help ensure that clustering in the high-dimensional feature space is more reliable. The comparison between Model-V3 and Model-V4 further demonstrates the effectiveness of extracting interaction patterns and integrating them into the prediction task. While most existing approaches have not quantitatively investigated social interactions, our study makes an initial attempt in this direction. Similarly, the comparison between Model-V3 and Model-V5 confirms the effectiveness of the MDN loss.

V. CONCLUSION

In this work, we proposed a latent-variable generative approach to interpret social interactions in pedestrian walking. The latent variable is learned from high-dimensional features of social interactions, while a maximum coding rate reduction criterion is employed to constrain the feature space and construct a social loss function. The clustering process is not hand-crafted but totally data-driven. Experimental results demonstrate the effectiveness of our method in capturing and explaining interaction patterns among pedestrians. For future work, we plan to apply our approach to a wider range of datasets and conduct more extensive experiments to further evaluate its generalization ability. We are also going to explore the number of clusters, i.e. the social interaction modes. Besides, interpretability inevitably reduces the accuracy of

prediction models. we will explore more effective ways of incorporating interaction patterns into a broader range of tasks.

REFERENCES

- [1] 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.
- [2] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezaatofghi, 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.
- [3] Inhwon Bae, Jean Oh, and Hae-Gon Jeon. Eigentrajectory: Low-rank descriptors for multi-modal trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10017–10029, 2023.
- [4] Francesco Marchetti, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. Smemo: social memory for trajectory forecasting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4410–4425, 2024.
- [5] Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection. *Neural networks*, 108:466–478, 2018.
- [6] Pu Zhang, Wanli Ouyang, Pengfei Zhang, Jianru Xue, and Nanning Zheng. Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12085–12094, 2019.
- [7] Xiaodan Shi, Xiaowei Shao, Zipei Fan, Renhe Jiang, Haoran Zhang, Zhiling Guo, Guangming Wu, Wei Yuan, and Ryosuke Shibasaki. Multimodal interaction-aware trajectory prediction in crowded space. In *AAAI*, pages 11982–11989, 2020.
- [8] Abdullallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcnn: 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*, pages 14424–14432, 2020.
- [9] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *European Conference on Computer Vision*, pages 507–523. Springer, 2020.
- [10] Yusheng Peng, Gaofeng Zhang, Xiangyu Li, and Liping Zheng. Stirnet: A spatial-temporal interaction-aware recursive network for human trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2285–2293, 2021.
- [11] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5308–5317, 2016.
- [12] Vasilij Karasev, Alper Ayvaci, Bernd Heisele, and Stefano Soatto. Intent-aware long-term prediction of pedestrian motion. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2543–2549. IEEE, 2016.
- [13] Lingyun Luke Li, Bin Yang, Ming Liang, Wenyuan Zeng, Mengye Ren, Sean Segal, and Raquel Urtasun. End-to-end contextual perception and prediction with interaction transformer. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5784–5791. IEEE, 2020.
- [14] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. *arXiv preprint arXiv:2103.14023*, 2021.
- [15] Yicheng Liu, Jinghuai Zhang, Liangji Fang, Qinrong Jiang, and Bolei Zhou. Multimodal motion prediction with stacked transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7577–7586, 2021.
- [16] 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.
- [17] Hao Cheng, Wentong Liao, Michael Ying Yang, Bodo Rosenhahn, and Monika Sester. Amenet: Attentive maps encoder network for trajectory prediction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 172:253–266, 2021.
- [18] Guangyi Chen, Junlong Li, Nuoxing Zhou, Liangliang Ren, and Jiwen Lu. Personalized trajectory prediction via distribution discrimination. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15580–15589, 2021.
- [19] Marion Neumeier, Andreas Tollkühn, Thomas Berberich, and Michael Botsch. Variational autoencoder-based vehicle trajectory prediction with an interpretable latent space. *arXiv preprint arXiv:2103.13726*, 2021.
- [20] Christopher M Bishop. Mixture density networks. 1994.
- [21] Osama Makansi, Eddy Ilg, Ozgun Cicek, and Thomas Brox. Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7144–7153, 2019.
- [22] Stuart Eiffert, Kunming Li, Mao Shan, Stewart Worrall, Salah Sukkarieh, and Eduardo Nebot. Probabilistic crowd gan: Multimodal pedestrian trajectory prediction using a graph vehicle-pedestrian attention network. *IEEE Robotics and Automation Letters*, 5(4):5026–5033, 2020.
- [23] Charlie Tang and Russ R Salakhutdinov. Multiple futures prediction. *Advances in Neural Information Processing Systems*, 32:15424–15434, 2019.
- [24] Lingyao Zhang, Po-Hsun Su, Jerrick Hoang, Galen Clark Haynes, and Micol Marchetti-Bowick. Map-adaptive goal-based trajectory prediction. *arXiv preprint arXiv:2009.04450*, 2020.
- [25] Junru Gu, Chen Sun, and Hang Zhao. Densetnt: End-to-end trajectory prediction from dense goal sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15303–15312, 2021.
- [26] He Zhao and Richard P Wildes. Where are you heading? dynamic trajectory prediction with expert goal examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7629–7638, 2021.
- [27] Harshayu Girase, Haiming Gang, Srikanth Malla, Jiachen Li, Akira Kanehara, Karttikeya Mangalam, and Chiho Choi. Loki: Long term and key intentions for trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9803–9812, 2021.
- [28] Jiachen Li, Fan Yang, Masayoshi Tomizuka, and Chiho Choi. Evolvegraph: Heterogeneous multi-agent multi-modal trajectory prediction with evolving interaction graphs. *ArXiv abs/2003.13924*, 2, 2020.
- [29] Patrick Dendorfer, Aljosa Osep, and Laura Leal-Taixé. Goal-gan: Multimodal trajectory prediction based on goal position estimation. In *Proceedings of the Asian Conference on Computer Vision*, 2020.
- [30] Kwan Ho Ryan Chan, Yaodong Yu, Chong You, Haozhi Qi, John Wright, and Yi Ma. Redunet: A white-box deep network from the principle of maximizing rate reduction. *Journal of machine learning research*, 23(114):1–103, 2022.
- [31] Yi Ma, Harm Derksen, Wei Hong, and John Wright. Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE transactions on pattern analysis and machine intelligence*, 29(9):1546–1562, 2007.
- [32] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, volume 26, pages 655–664. Wiley Online Library, 2007.
- [33] 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.
- [34] Julian N Marewski, Wolfgang Gaissmaier, and Gerd Gigerenzer. Good judgments do not require complex cognition. *Cognitive processing*, 11:103–121, 2010.
- [35] Mehdi Moussaïd, Dirk Helbing, and Guy Theraulaz. How simple rules determine pedestrian behavior and crowd disasters. *Proceedings of the National Academy of Sciences*, 108(17):6884–6888, 2011.
- [36] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezaatofghi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In *Advances in Neural Information Processing Systems*, pages 137–146, 2019.