Social-DPF: Socially Acceptable Distribution Prediction of Futures

Xiaodan Shi[1], Xiaowei Shao[1,2], Guangming Wu[1], Haoran Zhang[1], Zhiling Guo[1], Renhe Jiang[3],Ryosuke Shibasaki[1]

[1]Center for Spatial Information Science, the University of Tokyo
[2]Earth Observation Data Integration and Fusion Research Initiative, the University of Tokyo
[3]Information Technology Center, the University of Tokyo
{shixiaodan, jiangrh, zhang_ronan, guozhilingcc, huster-wgm, shiba}@csis.u-tokyo.ac.jp,
{shaoxw}@iis.u-tokyo.ac.jp

Abstract

We consider long-term path forecasting problems in crowds, where future sequence trajectories are generated given a short observation. Recent methods for this problem have focused on modeling social interactions and predicting multi-modal futures. However, it is not easy for machines to successfully consider social interactions, such as avoiding collisions while considering the uncertainty of futures under a highly interactive and dynamic scenario. In this paper, we propose a model that incorporates multiple interacting motion sequences jointly and predicts multi-modal socially acceptable distributions of futures. Specifically, we introduce a new aggregation mechanism for social interactions, which selectively models long-term inter-related dynamics between movements in a shared environment through a message passing mechanism. Moreover, we propose a loss function that not only accesses how accurate the estimated distributions of the futures are but also considers collision avoidance. We further utilize mixture density functions to describe the trajectories and learn multi-modality of future paths. Extensive experiments over several trajectory prediction benchmarks demonstrate that our method is able to forecast socially acceptable distributions in complex scenarios.

Introduction

The ability to predict long-term futures accurately lies at the heart of autonomous driving and social robots navigation(Kitani et al. 2012; Karasev et al. 2016; Liu et al. 2016; Lee et al. 2017; Su et al. 2017; Liang et al. 2019) where autonomous driving cars and social robots share the same ecosystem with humans. They adjust their paths by anticipating human movements, specifically, avoiding collisions or maintaining a safe distance from other people. Modeling human interactions is a challenging aspect of trajectory prediction task. Although humans can intuitively know how to interact with other people in crowds, it is not easy for machines to learn those interaction rules owing to the complexities and uncertainties of human crowds.

Since the success of recurrent neural network (RNN) on sequence modeling, RNN-based models have been well developed for use in trajectory prediction. Social LSTM, which introduces social pooling to calculate the global representations for interactions by aggregating the latent states of spatially proximal pedestrians, is an important development for real-world paths forecasting (Alahi et al. 2016). The existing research follow the way of Social LSTM but with improvements. The research (Huang et al. 2019; Mohamed et al. 2020; Li et al. 2020) utilize spatio-temporal graph representations to describe motion dynamics over time and space. By modeling the topography of graphs, the models can naturally model social interactions and the movement of people. By contrast, generative adversarial network (GAN)-based models are investigated to model the uncertainty of futures (Gupta et al. 2018; Sadeghian et al. 2019; Kosaraju et al. 2019). GAN-based models usually contains social mechanisms in the generators and forecast multiple plausible trajectories instead of a single future path. The distributions of futures can be plotted through samples from the generators. The existing methods usually predict a single future path or distributions of the futures by minimizing (or a combination of) L_2 loss, adversarial loss or maximizing the lower bound of the log-likelihood function between the distributions of estimated futures and the distributions of ground truth.

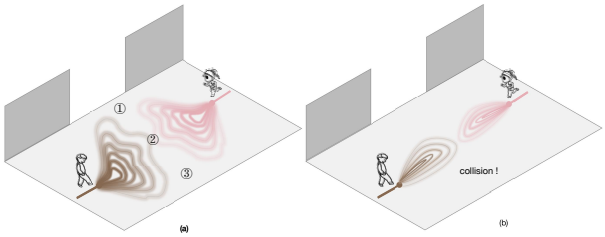

Figure 1: Objective of this study is to forecast socially acceptable distributions of futures. There are multiple plausible forthcoming paths in an interactive and dynamic scenario (for example, the boy walks to 1 and the girl walks to 2 or 3). The result of (b) is not socially acceptable owing to the collisions between them.

However, predicting socially acceptable trajectories still remains as an issue. First, the existing research usually use "pooling" as an aggregation mechanism which is limited to

model dynamic interactions and also leaky in information (Williams and Li 2018; Mohamed et al. 2020). They either use a Euclidean-distance based ordering structure to select a fixed number of neighbors or use Max/Average functions for pooling, where the former is not rational for use in dynamic real-world scenarios and the latter loses individual uniqueness (Kosaraju et al. 2019). Although it can be optimized by introducing attention mechanism (Luong, Pham, and Manning 2015; Fernando et al. 2018; Zhang et al. 2019), it still can't model interactions accurately(Sadeghian et al. 2019; Zhang et al. 2019). Second, although the existing research design architectures containing social mechanisms, their loss functions only access the accuracy of estimated distributions or locations are without considering social constraints, particularly collision avoidance, leading to a poor performance in terms of modeling social constraints.

To address the above limitations, we develop a trajectory prediction model named Social DPF for predicting socially acceptable multi-modal distributions of futures. To deal with dynamic social interactions, we propose a new aggregation mechanism by capturing the inter-related dynamics between multiple motion sequences over space and time. The new aggregation mechanism selectively incorporates the latent states of concurrent movements in a shared environment through two message passing gates and generates social states that reveals how people interact in crowds. To further generate socially plausible distributions of futures, we propose a loss function that not only accesses how the predictions meet the ground truth but also measures if the interactive futures collide. Training maximizes the lower bound on the log-likelihood of the data. Social-DPF further uses mixture density functions to describe human path and learns to model multi-modal futures jointly for all pedestrians.

Related Works

Social Compliant Trajectory Prediction

Since Social LSTM was proposed, many research have started investigating socially constrained RNN-based trajectory prediction models. Attention mechanisms were firstly utilized to improve Social LSTM by learning different weights of neighbors on an agent (Fernando et al. 2018; Sadeghian et al. 2019; Zhang et al. 2019). Some research embedded relative motion (e.g., the relative location and velocity) between pedestrians, followed by pooling the embeddings to generate a global feature for interactions, which is more intuitively than aggregating latent states of the RNNs directly (Becker et al. 2018; Shi et al. 2019). Instead of setting a neighborhood size, some research took into account all people in a scenario or followed a radial basis function to select a certain number of people(Kosaraju et al. 2019; Tang and Salakhutdinov 2019).

Spatio-Temporal Graphs for Trajectory Prediction

Structural RNN (Jain et al. 2016), combining high-level Spatio-termporal graphs (ST-graphs) with sequence modeling success of RNN, has made significant improvements on problem of human motion modeling. Hence, there are many research following this direction (Huang et al. 2019; Mohamed et al. 2020). ST-graphs for trajectory prediction are characterized with element points, spatial edges and temporal edges that represent pedestrians, social interactions and the temporal transition of trajectories respectively. ST-graphs make it easy to illustrate the topography of human motion and their interactions and provide a more direct and natural way to model long-term path forecasting by modeling different elements using RNNs.

Multi-Modal Trajectory Prediction

Human motions under crowded scenarios imply a multiplicity of modes. Given the observation, there are multiple plausible future paths. Some studies have combined GAN and RNNs to capture the uncertainty of long-term future paths (Gupta et al. 2018; Sadeghian et al. 2019; Amirian, Hayet, and Pettré 2019; Kosaraju et al. 2019). They usually contain an RNN-based encoder-decoder generator and an RNN-based decoder discriminator. Some research have applied Mixture Density Network (MDN) to map the distributions of future trajectories(Makansi et al. 2019; Shi et al. 2020). The article (Makansi et al. 2019), based on MDN, proposed a two-stage strategy that first predicted several samples of future with winner-take-all loss and then iteratively grouped the samples to multiple modes.

Loss Functions for Trajectory Prediction

The existing models can be classified as deterministic or stochastic models (Amirian, Hayet, and Pettré 2019). Some deterministic models utilizes L_2 loss function to predict a single path of future(Xue, Huynh, and Reynolds 2018; Becker et al. 2018). The others, such as MDN-based methods modeling the trajectories as a mixture model of bivariate Gaussian models, are trained by minimizing the negative log-likelihood function over the modes (Alahi et al. 2016; Shi et al. 2020; Makansi et al. 2019). The stochastic models, by contrast, are based on GAN and usually trained by minimizing the loss functions containing an L_2 loss and an adversarial loss L_{gan} (Gupta et al. 2018; Sadeghian et al. 2019). To truly learn multi-modality , Social BiGAT designed a loss function containing additional items to map the latent noise to an output trajectory (Kosaraju et al. 2019). However, the existing research have only accessed how accurate the estimated futures are. They haven't considered social interactions in loss functions.

Problem Formulation

We assume that each scenario is first preprocessed to obtain 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 any pedestrian i at any time instant t. The observation of pedestrian i is the past trajectory, which is represented as: $X_i^{1:\tau-1} = \{(x_i^t, y_i^t, u_i^t, v_i^t) | t = 1, 2, \cdots, \tau - 1\}$ while the future trajectory is $Y_i^{\tau:T} = \{(x_i^t, y_i^t) | t = \tau, \cdots, T\}$. We assume there are N agents in the scenario, $i = 1, 2, \cdots, N$.

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

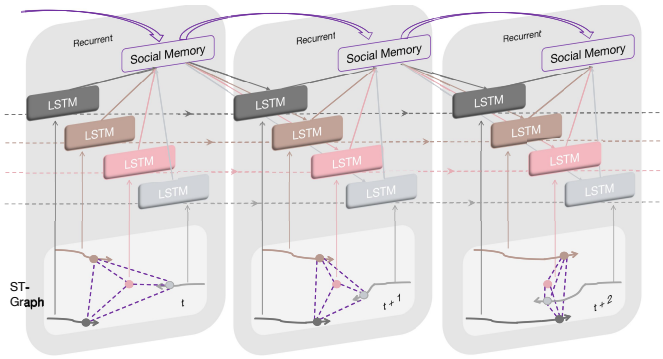

Figure 2: Overview of our model architecture. We utilize two LSTMs to capture the spatial and temporal cues, specifically one LSTM for single person's trajectories, and one LSTM for Social Memory which selectively integrates the latent states coming from the single person's LSTM. The details of Social Memory are illustrated in Fig.3.

Φ. Therefore, the distribution of futures is denoted as follows:

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

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

Methodology

Overall Architecture

Social DPF is an encoder-decoder model using two sets of LSTMs (LSTMs in the same set share weights) to represent the movement of agents and their interactions in a scene, as shown in Figure 2. Our model utilizes spatio-temporal graphs to represent human motions and their interactions. At any time instant, point elements of a graph are individuals characterized with location and velocity while the lines between two points are spatial edges representing their current interactions. We construct Social Memory, as depicted in Figure 3, which not only can capture the current interactions among people, it can also model how interactions change over time. The Social Memory takes use of LSTMs and are recurrent over time. The Social Memory takes hidden states from single person's LSTM as input and selectively integrate the inter-related hidden states through two gates to generate social states. Then social states storing interaction information are fed into single person's LSTM to generate latent features. Based on the latent features, our model directly outputs the parameters of the distributions of future trajectories through MDN combining a multilayer perception with Gaussian mixture models (GMMs). For the loss function, we use $\mathcal{L}_{mode}$ to estimate how closely the predicted distribution matches the distribution of target variables in the training data, whereas $\mathcal{L}_{collision1}$ and $\mathcal{L}_{collision2}$ achieve access if the estimated futures of agents collide with each other. Our loss function is based on Winner-Takes-All (WTA) loss which can prevent the model from collapsing into a single mode (Makansi et al. 2019).

Social Memory

The hidden states from single person's LSTM at time instant $t-1$ are represented as $\{h_i^{t-1}|i = 1, \cdots, N\}$ which are the input for Social Memory. We assume that a person with index i is the agent.

$$\tilde{h}_j^t = \psi_1(H_i^{t-1}, h_j^{t-1}; w_H^*) \quad (2)$$

where $\psi_1(\cdot)$ is the LSTM for Social Memory and its weights w_H^* are also shared among people in a scene, H_i^{t-1} is the social state of agent i at $t-1$ which stores the representation on how an agent interacts with other people, $j \in \{1, \cdots, N\}\setminus i$. To connect the current motions, we design a motion gate. The motion gate selectively obtains the motion information among people which have a genuine effect on the agent's path and contributes to a candidate social state.

$$\begin{aligned} a_j^{t-1} &= \phi_1(h_i^{t-1} \bigodot h_j^{t-1}; w_1^*) \\ \gamma_i^t &= \textstyle\sum_{j=1}^{N\setminus i} a_j^{t-1} \bigodot \tilde{h}_j^t \end{aligned} \quad (3)$$

where $\phi_1(\cdot)$ is a fully connected layer that connects an agent and other people, which are depicted as purple lines on ST-graph in Figure 2. Here, γ_i^t is the candidate social state from the motion gate, representing the interaction information between an agent and neighbors. We then construct output gate. The output gate learns the role of the neighbors in the agent's social interactions, which controls the extent to which social interactions remains in the social states.

$$\begin{aligned} g_j^{t-1} &= \phi_2(h_j^{t-1}; w_2^*) \\ o_i^{t-1} &= \textstyle\sum_{j=1}^{N\setminus i} g_j^{t-1} + \phi_3(H_i^{t-1}; w_3^*) \end{aligned} \quad (4)$$

where o_i^{t-1} is the feature from output gate, $\phi_2(\cdot)$ and $\phi_3(\cdot)$ are fully connected layers with $dropout = 0.50$, $\phi_2(\cdot)$ is with Sigmoid non-linearity, w_2^* and w_3^* are their weights respectively.

$$H_i^t = \gamma_j^t \bigodot o_i^{t-1} \quad (5)$$

where H_i^t is the social states which are then concatenated with the agent's current states for predicting the distributions of futures. The Social Memory is recurrent, thus H_i^t is also fed into the Social Memory in the next time step.

Path Forecasting

As mentioned in Section 3, the agent i at time instant t is characterized with location (x_i^t, y_i^t) and velocity (u_i^t, v_i^t). We embed them respectively to obtain the input for single person's LSTM.

$$f_i^t = [\phi_4((x_i^t, y_i^t); w_4^*), \phi_5((u_i^t, v_i^t); w_5^*)] \quad (6)$$

where $\phi_4(\cdot)$ and $\phi_5(\cdot)$ are fully connected layers with ReLU non-linearity and w_4^* and w_5^* are the embedding weights. We get the social states H_i^t of agent i at time t and concatenate it with f_i^t to predict the next state of the agent.

$$h_i^t = \psi_2(h_i^{t-1}, [f_i^t, H_i^t]; w_h^*) \quad (7)$$

Here, $\psi_2(\cdot)$ is single person's LSTM and its weights w_h^* are shared between all people in a scenario. To capture the multi-modality of future paths, we utilize MDN which combines a multilayer perception with GMMs. The next location of agent are given by the following:

$$\hat{y}_i^{t+1} \sim N_2(\alpha_i^t, \mu_i^t, \sigma_i^t) \tag{8}$$

where priors α_i^t, means μ_i^t and standard deviation σ_i^t are the output Gaussian mixture components of pedestrian i at time t. $\alpha_i^t = \{(\alpha_g)_i^t | g = 1, \ldots, M\}$, $\mu^t = \{(\mu_g)_i^t | g = 1, \ldots, M\}$, $\sigma_i^t = \{(\sigma_g)_i^t | g = 1, \ldots, M\}$ where $\mu_g = (\mu_x, \mu_y)$, $\sigma_g = (\sigma_x, \sigma_y)$ and M is the number of Gaussian models of MDN. Each Gaussian model is a bivariate Gaussian model. The probability density function of the next location conditioned on h_i^t is denoted as follows:

$$p(\hat{y}_i^{t+1} | h_i^t) = \alpha_i^t p(\hat{y}_i^{t+1} | \mu_i^t, \sigma_i^t) \tag{9}$$

We learn the mixing coefficients μ_i^t, σ_i^t and α_i^t through the network. To constrain α_i^t to lie within the range [0, 1] and to sum to unity, we use the $softmax$ function. Function $exp()$ is used to avoid standard deviation σ_i^t smaller than or equal to zero.

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

where $\{a_g | g = 1, \cdots, M\}$, $\{u_g | g = 1, \cdots, M\}$ and $\{z_g | g = 1, \cdots, M\}$is obtained through fully connected layers $\phi_\alpha(h_i^t)$, $\phi_\mu(h_i^t)$ and $\phi_\sigma(h_i^t)$ respectively.

Loss Function

To avoid collisions and truly learn the multi-modality of human motion, we design the loss function as indicated Eq.11, which combines three items for predicting the socially acceptable future trajectories for pedestrians.

$$\mathcal{L} = \mathcal{L}_{mode} + \lambda_1 \mathcal{L}_{collision1} + \lambda_2 \mathcal{L}_{collision2} \tag{11}$$

where $\mathcal{L}_{mode}$ is used to assess how accurate the estimated distributions are in predicting the future trajectories. In addition, $\mathcal{L}_{collision1}$ and $\mathcal{L}_{collision2}$ are introduced to prevent pedestrians colliding in crowds. To avoid collisions with others, the estimated distribution of the agent should first uncover the future locations of other people. We utilize $\mathcal{L}_{collision1}$ to capture this effect. Moreover, the estimated distributions at each time instant should be differ from each other. We use $\mathcal{L}_{collision2}$ to measure the different probability distributions of one person from others.

For $\mathcal{L}_{mode}$, instead of computing the negative log-likelihood function over all components of a mixture model, which easily collapses the model into a single mode, we always base the winner selection on the probability of the mixture model and multiply the winner probability with the learned weight as follows:

$$\mathcal{L}_{mode} = -\frac{1}{N} \sum_{t=\tau}^{T-1} \sum_{i=1}^{N} log(\alpha_i^t p(\hat{Y}_i^{t+1} | \mu_i^t, \sigma_i^t)) \tag{12}$$

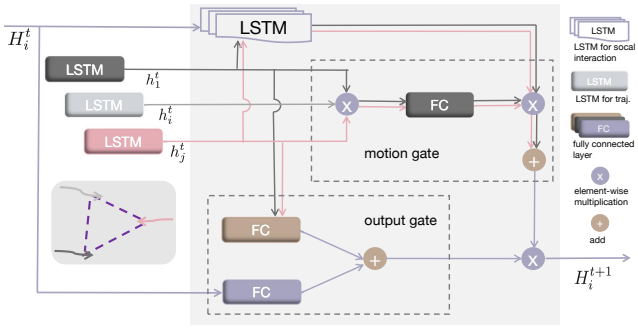

Figure 3: Illustration of Social Memory.

where, $\alpha_i^t = (\alpha_g)_i^t$, $\mu_i^t = (\mu_g)_i^t$, $\sigma_i^t = (\sigma_g)_i^t$ and $g = \arg\max_g p(\hat{Y}^{t+1} | \mu_g^t, \sigma_g^t)$. We assume that the predicted future path would be exactly the same as the ground truth and the neighbors are not supposed to lie within the estimated distribution of an agent. To achieve this effect, we introduced $\mathcal{L}_{collision1}$ as follows:

$$\begin{aligned} \mathcal{L}_{collision1} = -\frac{1}{N*(N-1)} \sum_{t=\tau}^{T-1} \sum_{i=1}^{N} \sum_{j=1}^{N \setminus i} \\ log(1 - \alpha_i^t p(Y_j^{t+1} | \mu_i^t, \sigma_i^t)) \end{aligned} \tag{13}$$

where, α_i^t, μ_i^t, σ_i^t are the same as in Eq.12. When neighbors don't lie within the probability distribution of an agent, $\mathcal{L}_{collision1}$ tends to be zero which has no effect on the entire loss function.

Here, $\mathcal{L}_{collision2}$ measures the amount of overlap between the predicted distributions of the pedestrians. If two predicted distributions can't be separate, the pedestrians tend to collide. We utilized Bhattacharyya distance to establish $\mathcal{L}_{collision2}$ as follows:

$$\begin{aligned} \mathcal{L}_{collision2} = \frac{1}{N*(N-1)} \sum_{t=\tau}^{T-1} \sum_{i=1}^{N} \sum_{j=1}^{N \setminus i} \\ log(\int_z \alpha_i^t \sqrt{q_i^t(z) q_j^t(z)} dz) \end{aligned} \tag{14}$$

where, $q_i^t(z) = N_2(z | \mu_i^t, \sigma_i^t)$, $q_j^t(z) = N_2(z | \mu_j^t, \sigma_j^t)$. A smaller $\mathcal{L}_{collision2}$ means that there is less overlap between two predicted distributions.

Experiments

In this section, the proposed model is evaluated on two publicly available datasets: UCY(Lerner, Chrysanthou, and Lischinski 2007) and ETH(Pellegrini et al. 2009). 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 at 2.5fps and transforming the coordinates of people into world coordinates in meters.

Implementation Details. Our model is trained end-to-end by minimizing the proposed loss function (Eq. 11). The experiments are implemented using Pytorch under Ubuntu

Method	Note	Evaluation (ADE(m)/FDE(m))					
		ETH-eth	ETH-hotel	UCY-univ	UCY-zara01	UCY-zara02	AVG
Linear	kalman filter	1.65/2.84	0.99/1.70	0.86/1.51	0.83/1.44	0.54/0.96	0.97/1.69
LSTM	offset is 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	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	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	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	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	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
S-DPF-$\mathcal{L}_{1,2,3}\mathcal{G}_1$	set1, $\mathcal{G}_1$,20 samples	0.89/1.16	0.53/0.95	0.82/1.30	0.49/0.74	0.59/0.90	0.66/1.01
S-DPF-$\mathcal{L}_{1,2,3}\mathcal{G}_2$	set1, $\mathcal{G}_2$,20 samples	0.76/0.96	1.24/1.26	0.94/1.32	0.88/1.12	0.81/1.01	0.93/1.13
S-DPF-$\mathcal{L}_1\mathcal{G}_{1,2}$	set2,$\mathcal{L}_1$,20 samples	0.71/0.95	0.35/0.54	0.62/0.78	0.40/0.65	0.45/0.57	0.51/0.70
S-DPF-$\mathcal{L}_{1,2}\mathcal{G}_{1,2}$	set2, $\mathcal{L}_{1,2}$,20 samples	0.66/0.94	0.42/0.58	0.59/0.74	0.37/0.64	0.41/0.51	0.49/0.68
S-DPF-$\mathcal{L}_{1,3}\mathcal{G}_{1,2}$	set2, $\mathcal{L}_{1,3}$,20 samples	**0.61/0.91**	0.40/0.54	0.57/0.75	0.36/0.63	0.39/0.49	0.47/0.66
S-DPF-$\mathcal{L}_{1,2,3}\mathcal{G}_{1,2}$-$\mathcal{V}_1$	set3, entire model,1 sample	0.69/1.35	0.39/0.84	0.61/1.00	0.40/0.89	0.39/0.84	0.50/0.98
S-DPF-$\mathcal{L}_{1,2,3}\mathcal{G}_{1,2}$-$\mathcal{V}_2$	set3, entire model,20 samples	0.66/0.92	**0.34/0.50**	0.50/**0.69**	0.34/0.59	0.32/**0.45**	**0.43/0.63**

Table 1: Quantitative results of baselines versus our method across datasets for predicting 12 future timesteps(4.8 sec) given 8 timesteps observations(3.2 sec)(lower is better). The results of Social LSTM, Social GAN are from (Gupta et al. 2018), the results of Sophie, Social BiGAT, Social STGCNN are from (Sadeghian et al. 2019; Kosaraju et al. 2019; Mohamed et al. 2020) respectively.

Datasets	Collision Evaluation (colliding persons per frame(%))				
	GT	Linear	S-GAN	S-STGCNN	S-DPF
ETH-eth	0.037	0.202	0.397	0.783	**0.030**
ETH-hotel	0.000	0.187	0.738	1.443	**0.012**
UCY-univ	0.304	2.408	2.519	3.741	**0.512**
UCY-zara01	0.000	0.523	0.186	1.100	**0.039**
UCY-zara02	0.044	0.833	0.904	3.100	**0.114**

Table 2: Average % of colliding people each frame in ETH&UCY. Two pedestrians are considered collided if their Euclidean distance is less than 0.2m.

16.04 LTS using a GTX 1080 GPU. The size of the hidden states of all LSTMs is set to 128. The embedding layers are composed of a fully connected layer with size 64 for Eq. 6 and 128 for the others. 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 a learning rate 0.001. We clip the gradients of the LSTM with a maximum threshold of 10 to stabilize the training process. We set λ_1 and λ_2 in Eq. 11 as 0.1. The model outputs GMMs with three components.

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 trajectory for 8 timesteps (3.2 sec) and show the 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 matrices.

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

2. Final displacement error (FDE): distance between prediction results and ground truth at the final timestep.

Baselines.The proposed model is compared with the following baselines.

1. Linear method. The second order Kalman Filter uses the position, velocity, acceleration.

2. LSTM. Human motion is modeled without considering human interactions. Offset is used as the input.

3.Social LSTM. It pools all hidden states of LSTMs for social interactions.

4. Social GAN. GAN-based model which considers social interactions and predicts multiple plausible futures.

5. Sophie. GAN-based model which considers both social and physical interactions to make more realistic predictions.

6. Social-BiGAT. This method uses a generator, local/global discriminators and a latent noise encoder to construct a reversible mapping between predicted paths and learned latent features of trajectories.

7.Social STGCNN: The method substitutes aggregation methods by modeling the interactions as a graph.

Ablation Study To describe how our model works, we also represent the results of various versions of our model Social DPF in an ablation setting using $\mathcal{L}_k\mathcal{G}_m$. Here, $\mathcal{L}_k$ signifies which loss the model is trained with (where $k = 1, 2, 3$ indicates $\mathcal{L}_{mode}$, $\mathcal{L}_{collision1}$ and $\mathcal{L}_{collision2}$, respectively.). In addition, $\mathcal{G}_m$ signifies which gate the model contains (where $m = 1, 2$ indicate the motion gate and output gate, respectively). For entire Social DPF, we also test two versions in a setting by:$\mathcal{V}_1$ uses the means of the distributions with maximum weights for testing; $\mathcal{V}_2$ draw 20 samples from the entire distributions for testing. We conducts three sets of ablation studies: set 1 containing $\mathcal{L}_{1,2,3}\mathcal{G}_m$ ($m = 1, 2$), set 2 containing $\mathcal{L}_k\mathcal{G}_{1,2}$ ($k = 1, 2, 3$), set 3 containing $\mathcal{L}_{1,2,3}\mathcal{G}_{1,2}$-$\mathcal{V}_n$($n = 1, 2$) that is the entire model of Social DPF.

Quantitative Evaluation

ETH and UCY. We compare our model to various baselines in Table 1, reporting the average displacement error (ADE) and final displacement error (FDE) for 12 timesteps of human movements. In general, the linear method performs worse than the other methods because it is limited to modeling the social context or multi-modality of human mo-

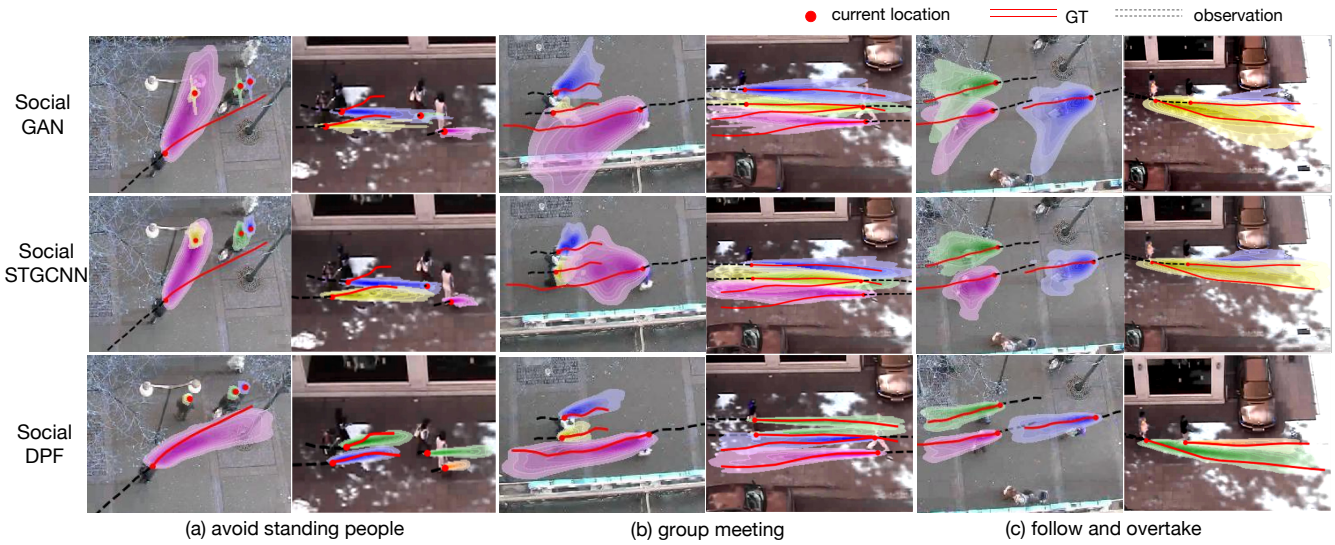

Figure 4: Comparison between Social GAN, Social STGCNN and our method over three sets of scenarios.

tion. Social LSTM only achieves an accuracy similar to that of LSTM, although it is trained with synthetic data and then fine-tuned on the benchmarks(Gupta et al. 2018). LSTM use offset as the input, which stabilizes the learning process and improves the performance. Sophie, Social GAN, Social Bi-GAT and Social STGCNN capturing the uncertainty of long-term movement all achieve better results than Social LSTM and basic LSTM.

The first set of our models tests how the motion gate and output gate perform with Social DPF. The models $\mathcal{L}_{1,2,3}\mathcal{G}_1$ and $\mathcal{L}_{1,2,3}\mathcal{G}_2$ solely modeling human interactions with motion gate and output gate respectively made a poor performance. By comparing the first and the third sets, we can easily find that motion gate and output gate together help Social DPF to better capture the long-term interactions among people. Interestingly, the motion gate seems to have a larger effect than output gate, which also reveals the importance of connecting inter-related dynamics. The second set of our models test the performance of each item of our loss function. As expected, $\mathcal{L}_{1,2}\mathcal{G}_{1,2}$ and $\mathcal{L}_{1,3}\mathcal{G}_{1,2}$ achieve better results than $\mathcal{L}_1\mathcal{G}_{1,2}$ on most of the datasets, which demonstrates that collision loss is able to help our model performs better. Interestingly, $\mathcal{L}_{1,3}\mathcal{G}_{1,2}$ performs slightly better than $\mathcal{L}_{1,2}\mathcal{G}_{1,2}$, which potentially implying $\mathcal{L}_{collision2}$ is more helpful than $\mathcal{L}_{collision1}$. On the other hand, $\mathcal{L}_1\mathcal{G}_{1,2}$ solely utilizing the proposed new aggregation mechanism for the social context performs well on average, which also demonstrates the ability of the proposed aggregation mechanism to model long-term social interactions. To further illustrate that collision loss can help to generate socially acceptable results, we also qualitatively compare the results of $\mathcal{L}_1\mathcal{G}_{1,2}$ and $\mathcal{L}_{1,2,3}\mathcal{G}_{1,2}$ in the next section (Qualitative Evaluation). The final model, $\mathcal{L}_{1,2,3}\mathcal{G}_{1,2}$, consisting of collision loss and two gates, outperforms the previous models, suggesting that combining both collision loss and new aggregation mechanism allows for robust predictions.

Collision Avoidance. To better understand our model's ability to produce socially acceptable futures, we also utilize another evaluation metric that reflects the percentage of near-collisions as in (Sadeghian et al. 2019). We consider two pedestrians to have collided if they come closer than 0.20m to each other. The average percentages of pedestrian collisions in ETH and UCY datasets are calculated as shown in Table 2 (average percentage of pedestrian collisions is the average of 20 samples from estimated distributions). Social DPF consistently outperforms other baselines in term of collision avoidance.

Qualitative Evaluation

To investigate the ability of social DPF to forecast socially acceptable futures distributions, we visualize three sets of scenarios from ETH-hotel and UCY-zara02 and compare the predictions of two state-of-the-art models, Social GAN and Social STGCNN, to that of our model (Fig.4). To compare their distributions more intuitively, we plot the entire distributions Social DPF predicted. In scenarios (a) where the walking pedestrians should adjust their courses to overtake people standing in front of them. Social GAN and Social STGCNN forecasted the futures distributions implying an incorrect walking direction or velocity, which further lead to collisions among people. Scenarios (b) depict groups meeting where collisions would also happen if pedestrians maintained their momentum. Social DPF jointly models the dynamics between movements and better aligns the predictions to social constraints than Social GAN and Social STGCNN. Scenarios (c) illustrate people following and overtaking in which Social DPF outperformed Social GAN and Social STGCNN by better predicting the walking speed, directions and avoidance behaviors.

To further illustrate that Social DPF is able to forecast multiple plausible distributions of futures, we also show three real-scenarios: overtaking, avoiding standing people,

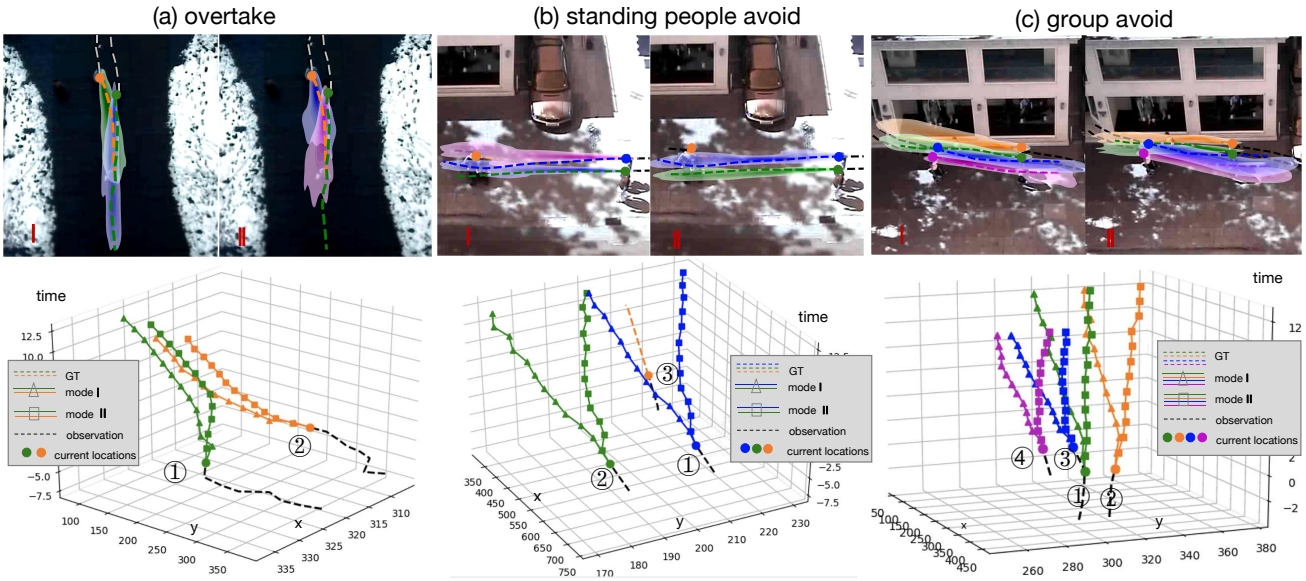

Figure 5: Multiple futures distributions that are socially acceptable. We show two sets of possible futures under each scenario where intense social interactions occur. The dynamics of people walking are shown in the 3D figures(we plot the average location of the distributions). Time is the z-axis and the same marker denote the same mode.

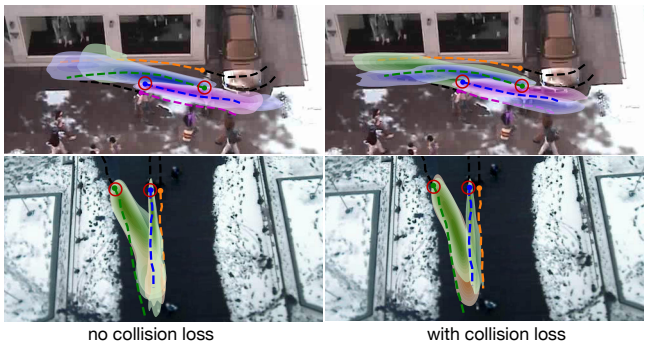

Figure 6: Comparison of collision loss.

and two-groups meeting where people have to alter their course to avoid collisions, as shown in Fig.5. In each scenario, two possible distributions (distributions of each agent with maximum and second maximum weights) of futures associated with velocity, walking direction are illustrated to show how people interact and navigate. In (a), the agent 1 will behave either in a mild way (I) or in an aggressive way (II) to overtake agent2. In (b), the agent 1 will collide with the standing pedestrian if he/she doesn't change walking direction. The agent 1 will overtake the standing people on the right or on the left. In (c), two groups meet in in front of Zara store, and the results show multiple plausible interactions between them. The pedestrians' distributions of the same mode seems to show a global coherency and conformity to social norms which also help Social DPF to predict socially acceptable results for pedestrians in a scene.

We also consider two real-world scenarios from ETH-eth and UCY-zara02 where intense social interactions occur to investigate how collision loss perform with Social DPF, as shown in Fig.6. In scenario1, two groups meet in a corner. Collision will happen if the agents doesn't change the walking course. In scenario2, the agent is merging into a group. Although Social DPF without collision loss is able to forecast the distributions of futures by adjusting walking direction or speed, $\mathcal{L}_{collision1}$ and $\mathcal{L}_{collision2}$prevent the agent from colliding with others which can help the model to predict socially acceptable results better.

Conclusion

We propose a trajectory prediction framework Social DPF, which jointly takes into account multiple interacting movements and predicts multi-modal socially acceptable distributions of futures. We introduce a novel aggregation mechanism called Social Memory to learn the long-term dynamic representations among pedestrians in a shared environment. Social Memory selectively integrates and stores the interaction information through two gates, motion gate and output gate. To better model the social constraints, we introduce collision loss to alleviate collision on futures distributions. Social DPF outperforms other state-of-the-art models over a number of publicly available datasets. We also demonstrate that it is able to provide more socially acceptable distributions by qualitatively analyzing the performance of Social DPF under scenarios such as group meeting, collision avoidance comparing to other baselines. Our model forecasting distributions of the same mode tend to show a global coherency and conformity to social norms. Future work will continue to explore it and extend our model to forecast all possible future modes of an interacting group. We also intend to consider multiple objects, such as bicycles, cars, and test the model performance with more benchmarks.

References

Alahi, A.; Goel, K.; Ramanathan, V.; Robicquet, A.; Fei-Fei, L.; and Savarese, S. 2016. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 961–971.

Amirian, J.; Hayet, J.-B.; and Pettré, J. 2019. Social ways: Learning multi-modal distributions of pedestrian trajectories with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Becker, S.; Hug, R.; Hübner, W.; and Arens, M. 2018. An evaluation of trajectory prediction approaches and notes on the trajnet benchmark. *arXiv preprint arXiv:1805.07663* .

Fernando, T.; Denman, S.; Sridharan, S.; and Fookes, C. 2018. Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection. *Neural networks* 108: 466–478.

Gupta, A.; Johnson, J.; Fei-Fei, L.; Savarese, S.; and Alahi, A. 2018. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2255–2264.

Huang, Y.; Bi, H.; Li, Z.; Mao, T.; and Wang, Z. 2019. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, 6272–6281.

Jain, A.; Zamir, A. R.; Savarese, S.; and Saxena, A. 2016. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 5308–5317.

Karasev, V.; Ayvaci, A.; Heisele, B.; and Soatto, S. 2016. Intent-aware long-term prediction of pedestrian motion. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2543–2549. IEEE.

Kitani, K. M.; Ziebart, B. D.; Bagnell, J. A.; and Hebert, M. 2012. Activity forecasting. In *European Conference on Computer Vision*, 201–214. Springer.

Kosaraju, V.; Sadeghian, A.; Martín-Martín, R.; Reid, I.; Rezatofighi, H.; and Savarese, S. 2019. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In *Advances in Neural Information Processing Systems*, 137–146.

Lee, N.; Choi, W.; Vernaza, P.; Choy, C. B.; Torr, P. H.; and Chandraker, M. 2017. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 336–345.

Lerner, A.; Chrysanthou, Y.; and Lischinski, D. 2007. Crowds by example. In *Computer graphics forum*, volume 26, 655–664. Wiley Online Library.

Li, J.; Ma, H.; Zhang, Z.; and Tomizuka, M. 2020. Social-wagdat: Interaction-aware trajectory prediction via wasserstein graph double-attention network. *arXiv preprint arXiv:2002.06241* .

Liang, J.; Jiang, L.; Niebles, J. C.; Hauptmann, A. G.; and 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*, 5725–5734.

Liu, Q.; Wu, S.; Wang, L.; and Tan, T. 2016. Predicting the next location: A recurrent model with spatial and temporal contexts. In *Thirtieth AAAI conference on artificial intelligence*.

Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* .

Makansi, O.; Ilg, E.; Cicek, O.; and Brox, T. 2019. 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*, 7144–7153.

Mohamed, A.; Qian, K.; Elhoseiny, M.; and Claudel, C. 2020. 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*, 14424–14432.

Pellegrini, S.; Ess, A.; Schindler, K.; and 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*, 261–268. IEEE.

Sadeghian, A.; Kosaraju, V.; Sadeghian, A.; Hirose, N.; Rezatofighi, H.; and 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*, 1349–1358.

Shi, X.; Shao, X.; Fan, Z.; Jiang, R.; Zhang, H.; Guo, Z.; Wu, G.; Yuan, W.; and Shibasaki, R. 2020. Multimodal Interaction-Aware Trajectory Prediction in Crowded Space. In *AAAI*, 11982–11989.

Shi, X.; Shao, X.; Guo, Z.; Wu, G.; Zhang, H.; and Shibasaki, R. 2019. Pedestrian trajectory prediction in extremely crowded scenarios. *Sensors* 19(5): 1223.

Su, H.; Zhu, J.; Dong, Y.; and Zhang, B. 2017. Forecast the Plausible Paths in Crowd Scenes. In *IJCAI*, volume 1, 2.

Tang, C.; and Salakhutdinov, R. R. 2019. Multiple futures prediction. In *Advances in Neural Information Processing Systems*, 15424–15434.

Williams, T.; and Li, R. 2018. Wavelet pooling for convolutional neural networks. In *International Conference on Learning Representations*.

Xue, H.; Huynh, D. Q.; and Reynolds, M. 2018. SS-LSTM: A hierarchical LSTM model for pedestrian trajectory prediction. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1186–1194. IEEE.

Zhang, P.; Ouyang, W.; Zhang, P.; Xue, J.; and Zheng, N. 2019. Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12085–12094.