LEARNING SOCIAL COMPLIANT MULTI-MODAL DISTRIBUTIONS OF HUMAN PATH IN CROWDS

Xiaodan Shi[1], Haoran Zhang[1],*Wei Yuan[1], Dou Huang[1], Zhiling Guo[1], Ryosuke Shibasaki[1]

[1]Center for Spatial Information Science, the University of Tokyo
{shixiaodan, zhang_ronan,miloyw, huangd, guozhilingcc, shiba}@csis.u-tokyo.ac.jp

Commission IV, WG IV/3

KEY WORDS: Pedestrian Trajectory Prediction, Social Interactions, Multi-modal, LSTM, Deep Learning.

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. It is not easy for prediction systems to successfully take into account social interactions and predict a distribution of future possible path in a highly interactive and dynamic circumstance. In this paper, we develop a data-driven model for long-term trajectory prediction, which naturally takes into account social interactions through a spatio-temporal graph representation and predicts multi-modes of future trajectories. Different from generative adversarial network (GAN) based models which generate samples and then provide distributions of samples, we use mixture density functions to describe human motion and intuitively map the distribution of future path with explicit densities. To prevent the model from collapsing into a single mode and truly capture the intrinsic multi-modality, we further use a Winner-Takes-All (WTA) loss instead of computing loss over all modes. Extensive experiments over several trajectory prediction benchmarks demonstrate that our method is able to capture the multi-modality of human motion and forecast the distributions of plausible futures in complex scenarios.

1. INTRODUCTION

Forecasting long-term future trajectories of dynamic pedestrians through crowded scenarios is of major importance for autonomous driving, social robots navigation and surveillance systems (Luber et al., 2010, Kitani et al., 2012, Karasev et al., 2016, Liu et al., 2016, Lee et al., 2017, Su et al., 2017). In autonomous driving and social robot navigation, autonomous driving cars and social robots share the same ecosystem with humans. They adjust their path by anticipating human movement, specifically, avoid collision or keep safe distance with other people. Predicting long-term trajectories in crowds is still a challenging topic due to the following properties of human motion.

1. Multi-modes of future trajectories. Given the observation of motion sequence, multiple future trajectory sequences are acceptable. It is more rational to map the distribution of future path instead of forecasting a single path especially for the task of long-term prediction.

2. Social interactions. Interactions between people happens frequently. Although humans can intuitively know how to interact with other people in crowds, it is not easy for machines to learn those rules due to complexity and dynamics of social interactions.

3. Scene context. Pedestrians motion should also obey physical constraints. Pedestrians walk on feasible terrains such as sidewalk or grass and avoid static obstacles such as roadblocks. Instead of encoding image of scenario into prediction model, physical constraints can also be learned from trajectories.

Since the success of recurrent neural network (RNN) on sequence modeling, the existing research focus on inventing RNN-based models for addressing the above problems and predicting long-term trajectories. Social LSTM (Alahi et al., 2016) pooling latent states coming from LSTMs of spatially proximal trajectories to model interactions, is a tipping point for real-world path forecasting. The existing research follow the direction of Social LSTM but with improvements. To directly model connections between people, articles (Gupta et al., 2018, Zhang et al., 2019) embed relative positions between agent and neighbors, then integrate those embeddings to generate a global feature for social interactions. Those methods learn interactions more intuitively by modeling relative motion between people, but they ignores time dependencies of long-term social interactions. Besides modeling social context, recent research pay more attentions to capture the intrinsic multi-modality of path forecasting. RNN-based generative adversarial network (GAN) are designed to capture uncertainty of future path (Gupta et al., 2018, Sadeghian et al., 2019, Kosaraju et al., 2019, Li et al., 2020a). Social GAN (Gupta et al., 2018) and Sophie (Sadeghian et al., 2019) utilize GAN-based (RNN-based generator and RNN-based discriminator) encoder-decoder architectures with social mechanism. They use generators to sample multiple future trajectories and plot distributions of those samples. We argue that it is not a direct way to model the multi-modes of trajectories and they don't yield complete distributions. They might only learn a single mode with high variance (Kosaraju et al., 2019).

To address the above limitations, we develop an encoder-decoder model which can learn the multi-modes of future trajectories. Different from GAN-based models which generate samples and then plot distributions of samples, we map the distribution of future path with explicit densities. To prevent the model from collapsing into a single mode and truly capture the intrinsic multi-modality, we use Winner-Takes-All (WTA) loss instead of computing loss over all modes. Besides, we utilize a spatio-temporal graph representation to naturally model social

* Corresponding author

interactions and ego-trajectories. We leverage relative motion between people while consider time dependencies of long-term social interactions. Instead of setting a certain neighborhood size or certain number of neighbors, we assume all people in a shared environment interacting and pool relative latent states between people through an attention mechanism. We test the model using classic trajectory prediction benchmarks and the experiments show promising results.

2. RELATED WORKS

2.1 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 (Alahi et al., 2016) but with improvements. Attention mechanism is introduced to learn neighbors' weights on agent (Zhang et al., 2019, Sadeghian et al., 2019, Fernando et al., 2018). 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 (Fernando et al., 2018). 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 (Zhang et al., 2019). Graph representation, specifically spatio-temporal graph (ST-graph) is well applied to illustrate human motion and their interactions (Karasev et al., 2016, Zhu et al., 2019, Shi et al., 2020, Mohamed et al., 2020, Yu et al., 2020, Peng et al., 2021). ST-graph provide a more direct and natural way to model interactions for trajectory prediction. Structure-RNN (Jain et al., 2016) combining high-level spatio-termporal 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 (Karasev et al., 2016). Social-STGCNN utilized spatio-temporal graph representation and proposed a weighted adjacency matrix to meansure the influence between pedestrians (Mohamed et al., 2020). Recently, Transformer is also used to model the motion and social interactions for trajectory prediction (Li et al., 2020c, Yuan et al., 2021, Liu et al., 2021). 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 (Li et al., 2020c). 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.

2.2 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 (Gupta et al., 2018, Sadeghian et al., 2019, Sohn et al., 2015, Cheng et al., 2021, Chen et al., 2021, Eiffert et al., 2020, Neumeier et al., 2021). Gupta A. et al. proposed Social GAN which contains RNN based encoder-decoder generator and RNN-based decoder discriminator (Gupta et al., 2018). 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 (Sadeghian et al., 2019). Some research apply Mixture Density Network (MDN) to map the distribution of future trajectories (Shi et al., 2020, Bishop, 1994, Makansi et al., 2019, Eiffert 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-Takes-All loss and then iteratively grouped the samples to multiple modes. There are also goal-based multi-trajectory prediction (Tang and Salakhutdinov, 2019, Mangalam et al., 2020, Li et al., 2020b, Zhang et al., 2020, Gu et al., 2021, Zhao and Wildes, 2021, Girase et al., 2021). 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 (Tang and Salakhutdinov, 2019, Li et al., 2020b). Another kind firstly forecasts multiple positional designations and then estimates futures matching the goal hypothesis (Dendorfer et al., 2020). 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.

3. 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: $X_i^{1:\tau} = \{(x_i^t, y_i^t, u_i^t, v_i^t) | t = 1, 2, \cdots, \tau\}$ while the future trajectories is $Y_i^{\tau:T} = \{(x_i^t, y_i^t) | t = \tau + 1, \cdots, T\}$.

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 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.

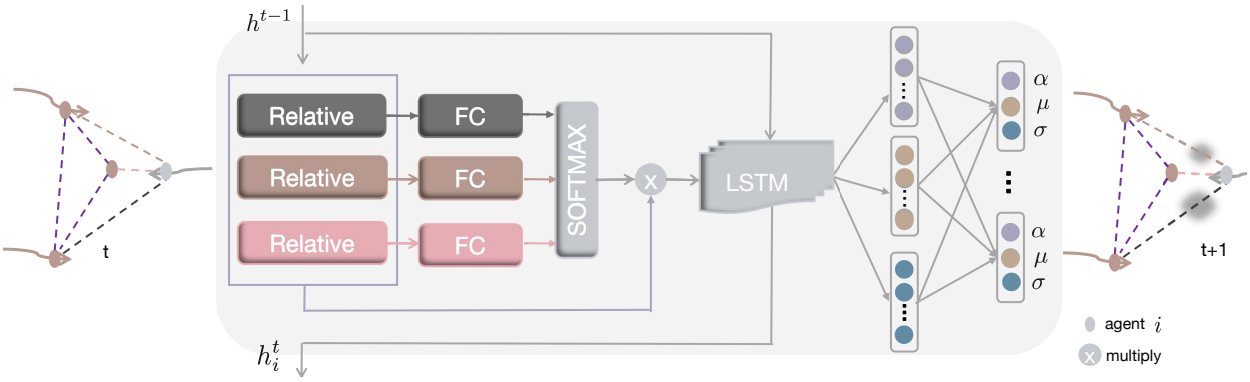

Figure 1. Overview of our method. We set an agent and illustrate the model. We use a separate LSTM for each person in a scene. The latent features from LSTMs are differencing to connect two interacting people, specially to construct spatial edges. The spatial edges are pooled to generate social features through an attention mechanism. Then the social features are fed into LSTM stacked with an Mixture Density Network (MDN) to forecast the distribution of future trajectories.

4. METHODOLOGY

4.1 Overall Architecture

Fig. 1 illustrates a unit of our encoder-decoder model at any time instance. We utilizes a spatial-temporal graph illustrated in Fig.2(a) to represent human motion and their interactions. At any time instance, point elements of a graph are people characterized with real-world location and velocity while lines between two points are spatial edges represent their current interaction. Temporal edges transferring adjacent graphs over temporal space. We leverage an LSTM to encode motion sequence. At any time instance, agent and neighbors are firstly connected by differencing latent features coming from last step as shown in Fig.2(b). Then an attention mechanism are used to pool all spatial edges to generate social features. The social features storing interaction information are be fed into LSTM to generate latent features which represent next states of people. Based on latent features, our model directly outputs the parameters of Gaussian Mixture Models (GMMs) which describe the distribution of future trajectories. It is easy to build a prediction model for any length based on the unit. In our encoder, social features are firstly concatenated with current state's embedding feature and then fed into LSTM. In decoder, social features are concatenated with latent features of last step.

4.2 Social Interactions

We assume person with index i is the agent. The hidden states from LSTM at time instance t-1 are represented as $\{h_j^{t-1}|j = 1, \cdots, N\}$ which are then utilized to construct spatial edges between agent and neighbors as $\mathcal{E}_j^t = \{h_j^{t-1} - h_i^{t-1}|j = 1, 2, \cdots, N\backslash i\}$. We assume all people in a shared scenario are allowed to interact. Most of the existing research construct spatial edges by embedding relative location between people. Unlike those research, we directly difference the latent states of LSTM between agent and neighbors, which also encourages the model to capture long time dependencies of social interactions. An attention mechanism is then applied to get the weights of neighbors on agent.

$$w_j^t = \frac{exp(\phi_1(\mathcal{E}_j^t; \omega_1^*))}{\sum_{k=1}^{N\backslash i}(exp(\phi_1(\mathcal{E}_k^t; \omega_1^*)))} \tag{2}$$

where $\phi_1(\cdot)$ is a fully connected layer for embedding spatial edges, ω_1^* are the embedding weights. w_j^t is a variable-length alignment vector, whose size equals the number of neighbors N-1.

$$s_i^t = \sum_{j=1}^{N\backslash i}(w_j^t * (h_j^{t-1} - h_i^{t-1})) \tag{3}$$

Given w_j^t as weights, the social feature s_i^t is computed as weighted sum over all spatial edges. s_i^t stores information how the agent interact with others. Some previous works use a Euclidean distance-based ordering structure to select a fixed number of neighbors, which is not rational in highly interacted, dynamic scenes. Some research used Max or Average functions to pool neighbors which may lost individual uniqueness. Here, we allow all neighbors to interact and selectively sum their features through an attention mechanism, which fits the realistic circumstance and doesn't lose individual uniqueness.

4.3 Path Forecasting

As mentioned in Section 3, the agent i at time instance 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 input for LSTM.

$$f_i^t = [\phi_2((x_i^t, y_i^t); w_2^*), \phi_3((u_i^t, v_i^t); w_3^*)] \tag{4}$$

where $\phi_2(\cdot)$ and $\phi_3(\cdot)$ are fully connected layers with ReLU non-linearity, w_2^* and w_3^* are the embedding weights. We represent social state for agent i at time t as s_i^t and concatenate it with f_i^{t-1} to predict next state of agent.

$$h_i^t = \psi_1(h_i^{t-1}, [f_i^t, s_i^t]; w_h^*) \tag{5}$$

$\psi_1(\cdot)$ is 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 that combines a multilayer perception with GMMs. The next location of agent conditioned on h_i^t are denoted as:

$$p(\hat{Y}_i^{t+1}|h_i^t) = \sum_{g=1}^{M} \alpha_g^t p(\hat{Y}_i^{t+1}|\mu_g^t, \sigma_g^t) \tag{6}$$

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

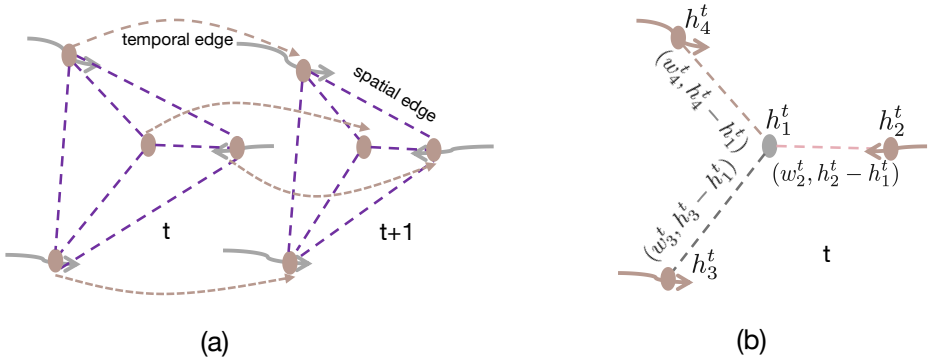

Figure 2. (a) spatio-temporal graph (b) spatial edges for agent

α_g^tthrough our network.

$$\begin{aligned} \alpha_g^t &= \frac{exp(a_g^t)}{\sum_{k=1}^{M} exp(a_k^t)} \\ \mu_g^t &= u_g^t \\ \sigma_g^t &= exp(z_g^t) \end{aligned} \tag{7}$$

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

4.4 Loss Function

The loss function is usually designed to compute negative log-likelihood of future trajectories over all components of a mixture model as Eq. (8). But it is easy to cause problems, such collapsing to a single mode.

$$\mathcal{L}_{all} = -\sum_{t=\tau}^{T-1} log(\sum_{g=1}^{M} \alpha_g^t p(\hat{Y}^{t+1}|\mu_g^t, \sigma_g^t)) \tag{8}$$

To capture the variety of multi-modes and truly learn the multi-modality of human motion, we design a WTA loss as Eq. (9). In the training process, we always base the winner selection on the probability. We compute loss by multiply the winner probability with learned weight. The weight of winner mode increase through training. Since our model is an encoder and a decoder model, it can be trained end to end and update all parameters.

$$\begin{aligned} \mathcal{L}_{wta} &= -\sum_{t=\tau}^{T-1} log(\alpha_g^t p(\hat{Y}^{t+1}|\mu_g^t, \sigma_g^t)) \\ g &= g\arg\max\ p(\hat{Y}^{t+1}|\mu_g^t, \sigma_g^t) \end{aligned} \tag{9}$$

5. EXPERIMENTS

In this section, the proposed model is evaluated on two publicly available datasets: UCY (Lerner et al., 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 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 1080 GPU. The size of hidden states of LSTM is set to 128. The embedding layers are composed of a fully connected layer with size 64 for Eq. (4) and 128 for 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 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.

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.

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

1. Linear. The second order Kalman Filter, which is modeled based on position, velocity, acceleration, is used as the linear method.

2. LSTM. Human motion is modeled without considering human interaction. Offset is used as input (Becker et al., 2018).

3. Social LSTM. This method models human interactions by pooling hidden states of spatially proximal motion sequences (Alahi et al., 2016).

4. Social GAN. This approach captures the multi-modality of future trajectory prediction, which contains a RNN based encoder-decoder generator and a RNN-based encoder discriminator. We consider one variant of Social GAN: best results of sampling 20 times (Gupta et al., 2018).

5. Sophie. This is a GAN-based model which takes into account both social and physical interactions to make more realistic predictions. We consider one variant of Sophie: best results of sampling 20 times (Sadeghian et al., 2019).

6. Social BiGAT. This method uses a generator, two discriminators (local discriminator and global discriminator) and a latent

Method	Year	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	2018,ECCV	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
S-LSTM	2016,CVPR	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	2018,CVPR	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
S-GAN	2018,CVPR	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
S-BiGAT	2019,NeurIPS	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
S-STGCNN	2020,CVPR	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
T-L$_{wat}V_1$	2021	not social,L$_{wat}$, $1sample$	0.72/1.25	1.10/2.09	0.70/1.38	0.50/1.10	0.37/0.80	0.68/1.32
S-T-L$_{all}V_1$	2021	L$_{all}$, $1sample$	0.63/1.12	1.44/2.76	0.53/1.10	0.41/0.90	0.31/0.69	0.66/1.31
S-T-L$_{wat}V_1$	2021	L$_{wat}$, $1sample$	0.64/1.20	0.87/1.65	0.65/1.21	0.50/1.08	0.33/0.75	0.60/1.18
T-L$_{wat}V_2$	2021	not social, L$_{wat}$, $20samples$	0.37/**0.59**	0.51/1.10	0.30/0.60	**0.21/0.45**	0.18/0.38	0.31/0.62
S-T-L$_{all}V_2$	2021	L$_{all}$, $20samples$	0.47/0.77	0.97/1.75	0.32/0.61	0.22/0.48	0.19/0.40	0.43/0.80
S-T-L$_{wat}V_2$	2021	L$_{wat}$, $20samples$	**0.36**/0.62	**0.45**/1.00	**0.27/0.59**	0.23/0.48	**0.17/0.36**	**0.30/0.61**

Table 1. Quantitative results of baselines vs. our method across datasets for predicting 12 future timesteps(4.8 sec) given 8 timesteps observation(3.2 sec). The results of S-LSTM, S-GAN are from (Gupta et al., 2018). The results of Sophie are from (Sadeghian et al., 2019). The results of S-BiGAT are from (Kosaraju et al., 2019). The results of S-STGCNN are from (Mohamed et al., 2020).Our model consistently outperforms other baselines (lower is better).

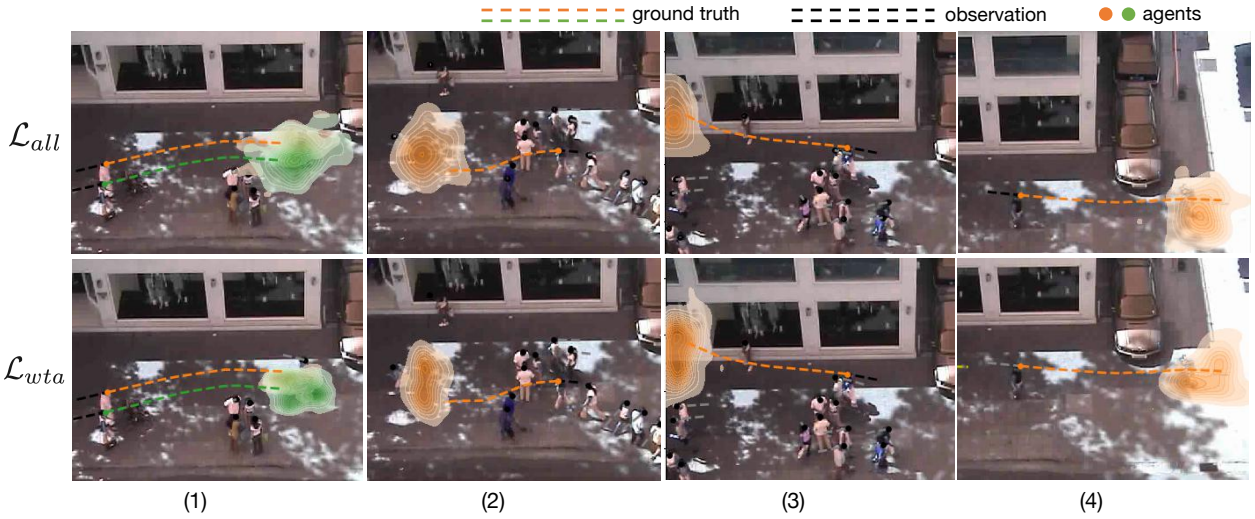

Figure 3. Distribution prediction of final step from two proposed models $S-T-\mathcal{L}_{all}V_2$ and $S-T-\mathcal{L}_{wta}V_2$ in four different sets. $\mathcal{L}_{all}$ is model with loss over all modes. $\mathcal{L}_{wta}$ is our model with WTA loss. (1)(2) show the distributions in social interactions. (3)(4) show the distributions in multiple entries/exits.

noise encoder to construct a reversible mapping between predicted paths and learned latent features of trajectories. We consider one variant of Social BiGAT: best results of sampling 20 times (Kosaraju et al., 2019).

Ablation Study. To explain how our model works, we also represent results of various versions of our models in an ablative setting by T-$\mathcal{L}_{wta}V_1/V_2$: with WTA loss and doesn't consider social interactions, S-T-$\mathcal{L}_{all}V_1/V_2$: with loss over all modes while considering social context, S-T-$\mathcal{L}_{wta}V_1/V_2$: our major model which use WTA loss and take into account social interactions. For all models, V_1 means sample once from mode with maximum weight while V_2 means sample 20 times from multi-modes.

5.1 Quantitative Evaluation

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 movement. In general, linear method performs worse than other methods because it is limited to model social context or multi-modality of human motion. Social LSTM only achieves similar accuracy as LSTM, although it is trained with synthetic data and then finetuned on benchmarks (Gupta et al., 2018). LSTM use offset as input, which makes the learning process stable and improves the performance. Sophie, Social GAN and Social BiGAT capturing the uncertainty of long-term movement achieve better results than other baselines.

Our first set of models $modelV_1$, which sample once from the mode with maximum weight, outperform baselines Linear, LSTM and Social LSTM. Even the first model $T-\mathcal{L}_{wta}V_1$ solely modeling pedestrian motion without considering social interactions achieves better performance than those three baselines. Our second set of models $modelV_2$ sample multiple times from multi-modes, make significant improvement than $modelV_1$ over two metrics. By comparing $S-T-\mathcal{L}_{wta}V_1$ and $T-\mathcal{L}_{wta}V_1$, we can tell modeling social context helps our model form better predictions in highly interactive scenarios. Interestingly, $T-\mathcal{L}_{wta}V_2$ achieves similar accuracy with $S-T-\mathcal{L}_{wta}V_2$ potentially suggesting the WTA loss is capable to truly forecast distributions of all possible path. The model $S-T-\mathcal{L}_{all}V_1$ trained by computing loss over all modes,

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-4-2022-91-2022 | © Author(s) 2022. CC BY 4.0 License.

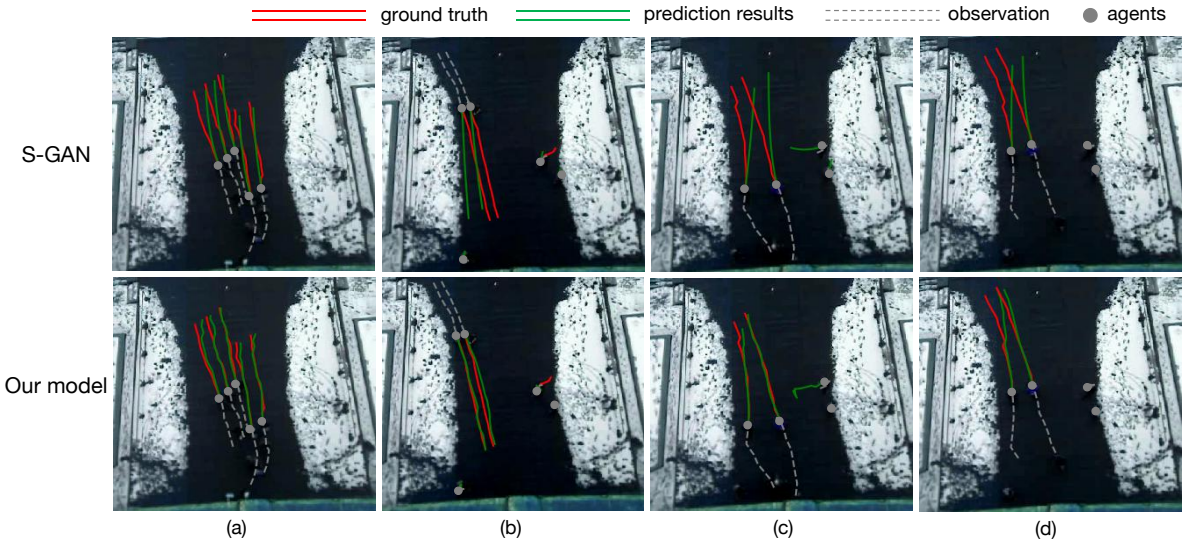

Figure 4. Comparison of predicted results from Social GAN (best result of 20 samples) and our model (best result of 20 samples).

achieves slightly better results than the model $S-T-\mathcal{L}_{wta}V_1$ over univ, zara01 and zara02. But the second version $S-T-\mathcal{L}_{all}V_2$ performs worse than $S-T-\mathcal{L}_{wta}V_2$, which indicates $\mathcal{L}_{wta}$ encourages the model forecast distributions covering more plausible futures than $\mathcal{L}_{all}$. Moreover, from the visualization in section 5.2, we also find model with $\mathcal{L}_{wta}$ can truly forecast the distributions of multi-modes while model with $\mathcal{L}_{all}$ tends to learn an average mode.

5.2 Qualitative Evaluation

Human motions tend to show multi-modality especially in long-term prediction or in a highly interactive circumstance. We further explore how our model performs by visualizing the predicted distributions of final step. Fig.3 show the visualization results of UCY-zara02 from two groups, social interactions in (1)(2) and multiple entries/exits in (3)(4). Warmer color indicates higher probability. We compare our model $S-T-\mathcal{L}_{all}V_2$ and $S-T-\mathcal{L}_{wta}V_2$. All the predicted spaces are socially and physically acceptable. From the visualization of distributions, we can see that model $S-T-\mathcal{L}_{wta}V_2$ better capture the multi-modality than $S-T-\mathcal{L}_{all}V_2$. In (1)(2), the agents adjust their path to avoid collision with neighbors in front. Except common sense rules like avoid collision, walking destinations and individual social manners can also have an effect on pedestrian's walking route. So the future trajectories tend to show multiple possibilities. Our model $S-T-\mathcal{L}_{wta}V_2$ successfully predict multiple plausible modes of future path. In (3), the agent might enter "zara" store or just walk along the road. Based on the observation, it is not easy to tell the agent's destination. The model $S-T-\mathcal{L}_{wta}V_2$ forecasts almost the same probabilities of entering "zara" store and going straight, which is more realistic than the result of $S-T-\mathcal{L}_{all}V_2$. The agents in (4) might also turn the corner instead of walking straight. The distribution from $S-T-\mathcal{L}_{wta}V_2$ clearly show multi-modes of future possible trajectories. Although model $S-T-\mathcal{L}_{all}V_2$ owns the capability of learning multi-modes but it tends to learn an average mode but loses the variety of multi-modes than $S-T-\mathcal{L}_{wta}V_2$.

We also visualize the results of Social GAN and our model $S-T-\mathcal{L}_{wta}V_2$ under the same scenarios to further investigate our model's performance in Fig.4. Both Social GAN and our model generate 20 samples and plot the best results. All the scenarios contain multiple interacting agents and multiple entries/exits. Based on the observations, the trajectories show the property of multi-modality. Our model outperforms Social GAN by better capturing multi-modes of future trajectories. In (b), the agents would avoid collision with the standing person in front, so they wouldn't behave as Social GAN predicts. In (a)(c)(d), people walk in a group and they might go straight, turn right or turn left. Our model generate results more accurately while Social GAN wrongly predict agents go straight instead of turning the corner.

6. CONCLUSION

We introduce an LSTM-based encoder-decoder model for long-term trajectory prediction in a highly interactive real-world circumstance. Unlike GAN-based models which firstly sample from generator and then derive distribution of future path from samples, we map distribution of human motion with explicit density by using Mixture Density Network. To avoid the model collapsing into a single mode and truly capture intrinsic multi-modality of human path, we further introduce Winner-Takes-All loss instead of computing loss over all modes. Besides, we assume all people in a shared environment are interacting and use an attention mechanism to sum all relative latent features between people to model social interaction. Finally, we show the efficacy of our method on several complicated real-life scenarios where social norms and multi-modality prediction must be followed.

REFERENCES

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

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

Bishop, C. M., 1994. Mixture density networks.

Chen, G., Li, J., Zhou, N., Ren, L., Lu, J., 2021. Personalized trajectory prediction via distribution discrimination. *Proceed-*

ings of the IEEE/CVF International Conference on Computer Vision, 15580–15589.

Cheng, H., Liao, W., Yang, M. Y., Rosenhahn, B., Sester, M., 2021. Amenet: Attentive maps encoder network for trajectory prediction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 172, 253–266.

Dendorfer, P., Osep, A., Leal-Taixé, L., 2020. Goal-gan: Multimodal trajectory prediction based on goal position estimation. *Proceedings of the Asian Conference on Computer Vision*.

Eiffert, S., Li, K., Shan, M., Worrall, S., Sukkarieh, S., Nebot, E., 2020. Probabilistic crowd GAN: Multimodal pedestrian trajectory prediction using a graph vehicle-pedestrian attention network. *IEEE Robotics and Automation Letters*, 5(4), 5026–5033.

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

Girase, H., Gang, H., Malla, S., Li, J., Kanehara, A., Mangalam, K., Choi, C., 2021. Loki: Long term and key intentions for trajectory prediction. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9803–9812.

Gu, J., Sun, C., Zhao, H., 2021. Densetnt: End-to-end trajectory prediction from dense goal sets. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15303–15312.

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

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

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

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

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

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

Lerner, A., Chrysanthou, Y., Lischinski, D., 2007. Crowds by example. *Computer graphics forum*, 26number 3, Wiley Online Library, 655–664.

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

Li, J., Yang, F., Tomizuka, M., Choi, C., 2020b. Evolvegraph: Heterogeneous multi-agent multi-modal trajectory prediction with evolving interaction graphs. *ArXiv, abs/2003.13924*, 2.

Li, L. L., Yang, B., Liang, M., Zeng, W., Ren, M., Segal, S., Urtasun, R., 2020c. End-to-end contextual perception and prediction with interaction transformer. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 5784–5791.

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

Liu, Y., Zhang, J., Fang, L., Jiang, Q., Zhou, B., 2021. Multimodal motion prediction with stacked transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7577–7586.

Luber, M., Stork, J. A., Tipaldi, G. D., Arras, K. O., 2010. People tracking with human motion predictions from social forces. *2010 IEEE International Conference on Robotics and Automation*, IEEE, 464–469.

Makansi, O., Ilg, E., Cicek, O., Brox, T., 2019. Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7144–7153.

Mangalam, K., Girase, H., Agarwal, S., Lee, K.-H., Adeli, E., Malik, J., Gaidon, A., 2020. It is not the journey but the destination: Endpoint conditioned trajectory prediction. *European Conference on Computer Vision*, Springer, 759–776.

Mohamed, A., Qian, K., Elhoseiny, M., Claudel, C., 2020. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14424–14432.

Neumeier, M., Tollkühn, A., Berberich, T., Botsch, M., 2021. Variational Autoencoder-Based Vehicle Trajectory Prediction with an Interpretable Latent Space. *arXiv preprint arXiv:2103.13726*.

Pellegrini, S., Ess, A., Schindler, K., Van Gool, L., 2009. You'll never walk alone: Modeling social behavior for multi-target tracking. *2009 IEEE 12th International Conference on Computer Vision*, IEEE, 261–268.

Peng, Y., Zhang, G., Li, X., Zheng, L., 2021. Stirnet: A spatial-temporal interaction-aware recursive network for human trajectory prediction. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2285–2293.

Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Rezatofighi, H., Savarese, S., 2019. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. *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., Shibasaki, R., 2020. Multimodal interaction-aware trajectory prediction in crowded space. *AAAI*, 11982–11989.

Sohn, K., Lee, H., Yan, X., 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 3483–3491.

Su, H., Zhu, J., Dong, Y., Zhang, B., 2017. Forecast the plausible paths in crowd scenes. *IJCAI*, 1, 2.

Tang, C., Salakhutdinov, R. R., 2019. Multiple futures prediction. *Advances in Neural Information Processing Systems*, 32, 15424–15434.

Yu, C., Ma, X., Ren, J., Zhao, H., Yi, S., 2020. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. *European Conference on Computer Vision*, Springer, 507–523.

Yuan, Y., Weng, X., Ou, Y., Kitani, K., 2021. AgentFormer: Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting. *arXiv preprint arXiv:2103.14023*.

Zhang, L., Su, P.-H., Hoang, J., Haynes, G. C., Marchetti-Bowick, M., 2020. Map-Adaptive Goal-Based Trajectory Prediction. *arXiv preprint arXiv:2009.04450*.

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

Zhao, H., Wildes, R. P., 2021. Where are you heading? dynamic trajectory prediction with expert goal examples. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7629–7638.

Zhu, Y., Qian, D., Ren, D., Xia, H., 2019. Starnet: Pedestrian trajectory prediction using deep neural network in star topology. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 8075–8080.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-4-2022-91-2022 | © Author(s) 2022. CC BY 4.0 License.