STIRNet: A Spatial-temporal Interaction-aware Recursive Network for Human Trajectory Prediction

Yusheng Peng[1], Gaofeng Zhang[2], Xiangyu Li[1], Liping Zheng[1,2] *
[1]School of Computer Science and Information Engineering, Hefei University of Technology
[2]School of Software, Hefei University of Technology

Abstract

Pedestrian trajectory prediction is one of the important research topics in the field of computer vision and a key technology of autonomous driving system. However, it's full of challenges due to the uncertainties of crowd motions and complex interactions among pedestrians. We propose a Spatio-temporal Interaction-aware Recursive Network (STIRNet) to predict multiply socially acceptable trajectories of pedestrians. In this paper, a recursive structure is used to capture spatio-temporal interactions by spatial modeling and temporal modeling alternately. At each time-step, the spatial interactions are modeled by a graph attention network, in which the nodes feature are represented by temporal motion features. The learned spatial interaction context is used to capture temporal motion features through an LSTM model. The temporal motion features are used to infer future positions and update nodes features. Experimental results on two public pedestrian trajectory datasets (ETH and UCY) demonstrate that our proposed model achieves superior performances compared with state-of-the-art methods on ADE and FDE metrics.

1. Introduction

Pedestrian trajectory prediction is of major significance in several applications such as autonomous driving, robot navigation, and surveillance systems. For example, in surveillance systems, forecasting pedestrian trajectories is critical in helping identify suspicious activities [23, 20].

In recent years, with the development of deep learning, the deep neural networks including LSTM, GAN are widely used in pedestrian trajectory prediction and achieve great success. In such deep learning prediction methods, pooling mechanisms [2, 10, 3], attention mechanisms [8, 31, 22] and graph neural network mechanisms [14, 18, 39] are often used to model the complex and subtle social interaction among pedestrians. In the view that pedestrians have different impacts on each other, some of the pooling mechanisms and graph neural network mechanisms incorporate attention mechanisms to model social interactions.

However, most of the models focus on modeling spatial interactions among pedestrians. Xu *et al*. [34] design a spatio-temporal attention module to model the spatio-temporal interactions among pedestrians. In contrast, STGAT [14] and AST-GNN [39] models model spatial interactions firstly and then feed the spatial interaction contexts to the temporal model to capture the spatio-temporal interaction features. Inspired by these works, we adopt a novel recursive structured network via graph attention network and LSTM to model spatio-temporal interactions.

In this paper, we propose a Spatio-temporal Interaction-aware Recursive Network (STIRNet) for pedestrian trajectory prediction. A GAT is adopted to model spatial interactions among pedestrians at each time-step, where the nodes features are represented by temporal motion features. Besides, the output spatial interaction contexts of GAT are fed to the LSTMs to capture temporal motion features. The learned motion features are used to infer future positions and update nodes features at next time-step.

The main contributions of this paper are summarized as follows:

- We propose a novel recursive structured trajectory prediction model which can capture spatio-temporal interactions by alternately performing temporal modelling and spatial modelling.
- A GAT is adopted to model spatial interactions from nodes features which are represented by temporal motion features.
- Experiments on ETH and UCY datasets show that STIRNet significantly improves pedestrian trajectory prediction and achieving state-of-the-art performance on two popular benchmarks.

2. Related Work

[*] Corresponding author: Liping Zheng (e-mail: zhenglp@hfut.edu.cn).

2.1. Trajectory Prediction Methods

Forecasting human trajectory has been researched for decades. In the early stages, many classic approaches are applied such as linear regression and Kalman filter [15], Gaussian processes [6] and Markov decision processing [17]. However, it is hard to model complex social interactions and normally fail in crowded scenes via these methods.

In recent years, the Long Short-Term Memory (LSTM) model has achieved great success in various sequence prediction tasks and is widely used in pedestrian trajectory prediction methods [2, 38, 35]. Gupta *et al*. [10] introduce Generative Adversarial Networks (GANs) to trajectory prediction task and propose a variety loss function that encourages the network to produce socially plausible trajectories. Inspired by this, a large number of GAN-based trajectory prediction models [5, 3, 25] emerged later. As another popular generative model, Conditional Variational Auto-Encoder (CVAE) is also adopted in various trajectory prediction methods [16, 24, 37]. As the Temporal Convolutional Network (TCN) reached or even exceeded the Recurrent Neural Network (RNN) in multiple tasks, some scholars use TCN to replace the RNN model and achieve success in pedestrian trajectory prediction [32, 30].

In this paper, we introduce a novel recursive structure network to predict trajectory where the LSTM is used for temporal modeling to capture motion features. Besides, a VAE model is employed during the training stage to encourage the proposed model to generate multimodal socially plausible trajectories.

2.2. Interactions Modeling in Trajectory Prediction

As a pioneering work, the social force model [11] achieves great success in interaction modeling and is widely used in crowd analysis and robotics. Social force models work well on interaction modeling while performing poorly on trajectory prediction [1]. Recently, in deep learning-based models, pooling mechanisms [2, 26, 27] that approximate crowd interaction are widely used. Besides, recent works consider pedestrians as nodes in a graph and intergrate information of the proximal pedestrians with attention mechanisms [29, 36, 39]. Explicit message passing allows the network to model more complex social behaviors.

The methods mentioned above focus on modeling spatial interactions between pedestrians. However, some scholars propose predicting trajectories by modeling spatio-temporal interaction among pedestrians. The existing methods [22, 34, 14, 39] are based on the seq2seq structure and capture spatio-temporal interaction context explicitly. Different from them, the proposed STIRNet model is based on a recursive structure and learns the latent motion feature which contains both spatial and temporal contexts by performing spatial and temporal modeling alternately. Specially, in STGAT [14], the GAT is used to capture spatial interaction contexts of historical trajectories, and the contexts are treated for temporal modeling through a LSTM to acquire the spatial-temporal context. Similar to STGAT, we adopt GAT and LSTM for spatial modeling and temporal modeling respectively. Innovatively, the GAT and LSTM are executed alternately through a recursive structure that allows the temporal and spatial contexts to be fully integrated.

2.3. Graph Neural Networks in Trajectory Prediction

Graph Neural Networks (GNNs) are powerful deep learning architectures for processing graph-structured data. In the pedestrian trajectory prediction task, pedestrians in the scene can be treated as nodes in the graph. In these works [28, 4], Graph Convolutional Networks (GCNs) are used as message passing schemes to aggregate social information from adjacency nodes. In particular, Graph Attention Networks (GATs) implement efficient weighted message passing between nodes and achieve great success in trajectory prediction [14, 18, 39].

In these methods, the GNNs are often used to model spatial interactions. In this paper, we also adopt GAT to model spatial interactions, in which the nodes features are represented by temporal features.

3. Proposed Method

The overview of the STIRNet model is illustrated in Figure 1. A recursive framework is adopted in the STIRNet model. For each time-step, the encoders embed the positions to high-dimensional features and the decoders are designed for inferring future positions from high-dimensional features. The GAT is employed to model spatial interactions from nodes features. Then the spatial interaction context is coupled with the encoding from the encoder and fed to the LSTM to capture motion feature. Besides, we design a VAE-based latent variable generator to generate latent variables in the training stage to encourage the model to predict multiply socially acceptable positions in the test stage.

3.1. Problem Formulation

The trajectory prediction task is formulated as one that estimates the positions of all pedestrians in the scene in the future period of time from their history trajectories. We assume that there are N pedestrians involved in the scene. Given certain observed positions $\{p_i^t|(x_i^t, y_i^t), t = 1, 2, ..., T_{obs}\}$ of pedestrians *i* of T_{obs} time-steps, our goal is predicting the positions $\{p_i^{t'}|(\widehat{x}_i^{t'}, \widehat{y}_i^{t'}), t' = T_{obs}+1, T_{obs}+2, ..., T_{pred}\}$ of future T_{pred} time-steps.

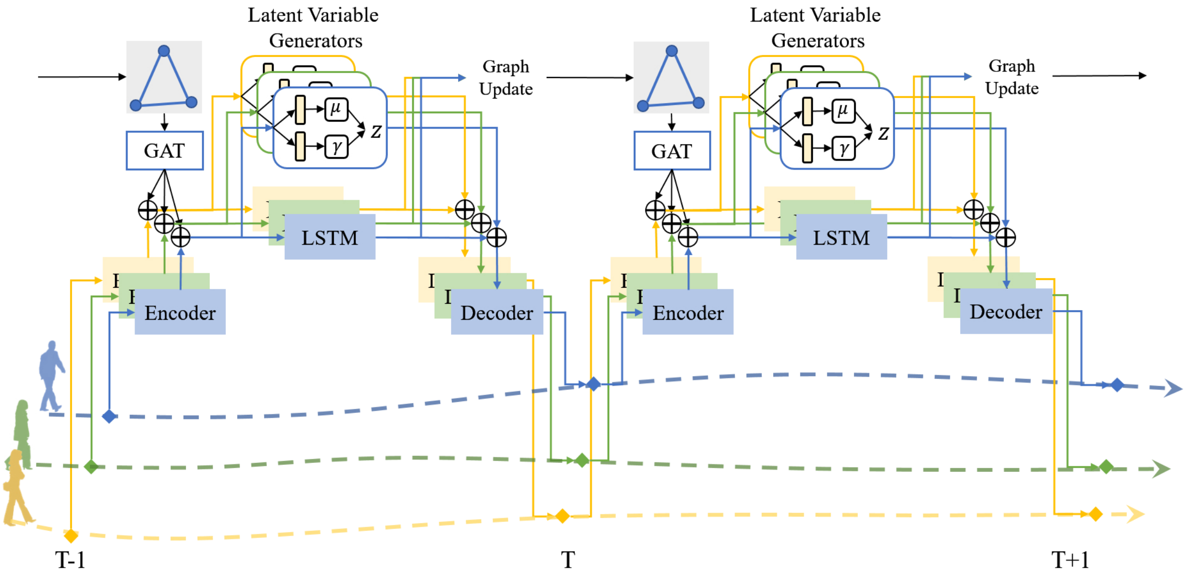


Figure 1. The architecture of the proposed STIRNet model. The model is built with a recursive structure. At each time-step, the encoders embed the positions to high-dimensional features, and the decoders decode the high-dimensional feature into position coordinates of the next time-step. The GAT and LSTM are employed for spatial modeling and temporal modeling to capture spatial interaction context and motion feature. The spatial modeling and temporal modeling are performed in an alternate manner to recursively predict future positions. Besides, a VAE is used to generate latent variables in training stage for multimodal trajectory prediction.

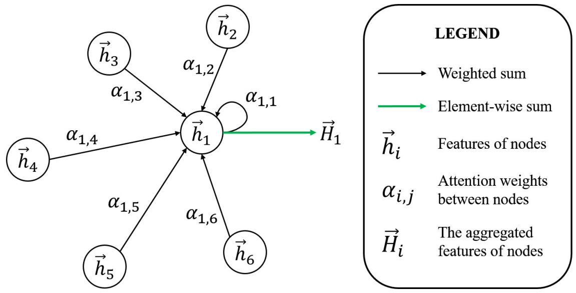


Figure 2. An illustration of graph attention network. GAT assigns different weights to different neighborhood nodes and aggregates features of them.

3.2. GAT-based Spatial Interaction Modeling

The social interactions that exist between pedestrians are crucial for pedestrian trajectory prediction. In the deep learning-based trajectory prediction models, pooling mechanisms, attention mechanisms are often used to model the social interaction among pedestrians. Besides, GNNs are also widely adopted in social interaction modeling. As one of the popular GNNs, GAT adopts the attention mechanism and can assign different weights to different nodes. In view of the success of GAT in pedestrian prediction [14, 18, 39], we adopt GAT to model the social interaction among pedestrians.

In this section, the pedestrians in the scene are treated as nodes in the graph, and the node features are represented by the temporal motion features of pedestrians. The GAT mechanism we adopted is illustrated in Fig. 2. For each time-step t, the input of GAT is a set of features of nodes which is represented by $h = \left\{\vec{h}_i^{t-1} | \vec{h}_i^{t-1} \in \mathbb{R}^F, \forall i \in \{1, \ldots, N\}\right\}$, N is the number of nodes, and F is the feature dimension of each node. Firstly, the nodes features are transformed to distinct intermediate representations through a learnable linear transformation $\mathbf{W} \in \mathbb{R}^{F \times F'}$. Then, the self-attention mechanism is performed on these nodes, and the coefficient of the node pair (i, j) is computed by:

$$\alpha_{i,j}^t = \frac{\exp\left(LR\left(a^T\left[\mathbf{W}\vec{h}_i^{t-1} \parallel \mathbf{W}\vec{h}_j^{t-1}\right]\right)\right)}{\sum_{k\in\mathbb{N}_i}\exp\left(LR\left(a^T\left[\mathbf{W}\vec{h}_i^{t-1} \parallel \mathbf{W}\vec{h}_k^{t-1}\right]\right)\right)} \qquad (1)$$

where $\alpha_{i,j}^t$ represents the impact of node j on node i at time-step t, and $\mathbb{N}_i$ represents the neighbors of node i on the graph. LR is a LeakyReLU function, $\|$ is the concatenation operation, $a \in 2F'$ is a learnable weight vector, and $.^T$ represents transposition.

After getting the attention weights among nodes, the features of neighboring nodes are aggregated by an aggregate function. For instance, the aggregated feature of node i at time-step t is given by:

$$\vec{H}_i^t = \sum_{k\in\mathbb{N}_i} \alpha_{i,j}^t \mathbf{W}\vec{h}_j^{t-1} \qquad (2)$$

where $\vec{H}_i^t$ is the aggregated hidden state for pedestrian i at time-step t, which contains the spatial influence from other pedestrians. And it is regarded as the spatial interaction context of pedestrian i at the current time-step.

3.3. LSTM-based Temporal Modeling

Pedestrian trajectory prediction is a sequential prediction task which predicts the future trajectory through the observed pedestrian trajectory. Most pedestrian trajectory prediction works [10, 3, 14, 13, 7] adopt the sequence-to-sequence architecture, in which the observed trajectory is treated as input sequence to feed to encoder, and the encoding feature is decoded to infer the future trajectory that can be treated as the output sequence. Instead, a small number of models [2, 38, 35] are based on recursive structures, which perform the same operation at each time-step to extract features and infer the position of the next time-step. In this paper, we propose a novel pedestrian trajectory prediction model via recursive structure.

As a popular variant of recurrent neural network, LSTM achieves great success in sequence tasks [12, 9]. In recursive structure trajectory prediction models [2, 38, 35], LSTMs are adopted as main bodies to capture motion features, and the hidden state of LSTM at each time-step is used to infer the position of next time-step. In our proposed model, the LSTM is used for temporal modeling to capture temporal motion features from spatial interaction context. Specifically, at each time-step, the positions of pedestrians are encoding to high dimensional features through encoders firstly. Then, the spatial interaction contexts acquired from GAT in Sect. 3.2 and encodings are coupled and served as the current inputs of LSTMs. The hidden states of LSTMs are regarded as motion features captured at current time-step to infer the positions of the next time-step. In addition, the hidden states are used to update the nodes features of the graph, which means that the hidden states are treated as nodes features of the graph of the next time-step. The key operations are formulated by:

$$e_i^t = \phi(x_i^t - x_i^{T_{obs}}, y_i^t - y_i^{T_{obs}}; W_e) \quad (3)$$

$$h_i^t = \text{LSTM}(h_i^{t-1}, \vec{H}_i^t, e_i^t; W_l) \quad (4)$$

where $\phi(\cdot)$ is an embedding function with ReLU nonlinearity of encoder, W_e is the embedding weight. The LSTM weight is denoted by W_l. These parameters are shared among all the pedestrians in the whole scene. H_i^t is the social interaction context which is the output of the GAT in Sect. 3.2. Similar to [38], we use the normalized absolute(Nabs) position which shifts the origin to the latest observed time slot.

After spatial modeling and temporal modeling, the learned hidden state h_i^t contains spatial context and temporal context implicitly. In the recursive structure, the hidden state h_i^t is treated as node feature at the next time-step to update graph node information for spatial modeling. By this alternate manner, the spatial contexts and temporal contexts are fully integrated into the learned latent motion features, so as to better inferring the future positions.

3.4. Latent Variable Generator

For multimodal prediction, a VAE-based latent variable generator is designed to generate the latent variables μ and γ (see Fig. 1). The existing trajectory prediction methods use latent variables to handle multimodality where the latent variables are directly sampled from the normal distribution [10, 14] or a multivariate normal distribution conditioned on the observed trajectories [19, 21]. To make our latent variables more aware of social cues, we design a novel VAE model to learn the parameters of the sampling distribution from the current position and spatial interaction context. To this end, the concatenated feature $e_i^t \oplus \vec{H}_i^t$ is passed to two different fully connected layers to yield the mean vector μ_i and logarithmic variance γ_i and finally z_i for the downstream decoder:

$$\mu_i = \phi_\mu(e_i^t \oplus \vec{H}_i^t; W_\mu) \quad (5)$$

$$\log \delta_i^2 \triangleq \gamma_i = \phi_\mu(e_i^t \oplus \vec{H}_i^t; W_\delta) \quad (6)$$

$$z_i \sim \mathcal{N}(\mu_i, diag(\delta_i^2)) \quad (7)$$

where W_μ and W_δ are trainable weights of ϕ_μ and ϕ_δ. The re-parameterization trick [19] is applied to sample the latent variable z_i.

3.5. Trajectory Prediction

To infer the position of next time-step, the concatenated feature $h_i^t \oplus z_i$ at time *t* is fed to the decoder:

$$[\Delta\hat{x}_i^{t+1}, \Delta\hat{y}_i^{t+1}]^T = W_p[h_i^t \oplus z_i] \quad (8)$$

$$(\hat{x}_i^{t+1}, \hat{y}_i^{t+1}) = (\Delta\hat{x}_i^{t+1} + x_i^{T_{obs}}, \Delta\hat{y}_i^{t+1} + y_i^{T_{obs}}) \quad (9)$$

where W_p is a weight matrix. In the test stage, we can sample z from $\mathcal{N}(0, 1)$ multiple times to generate multiple future positions.

3.6. Implementation Details

The parameters of the STIRNet model are directly learned by minimizing the $L2$ loss between the predicted positions and ground truth. **W** in GAT is of shape 64×64. The dimension of encoder vector e_i^t in Eq.3 is set to 32, and the dimension of hidden states of LSTM cells is set to 64. The dimension of latent variable z is set to 16. All trajectory segments in the same time window are regarded as a mini-batch, as they are processed in parallel. Adam optimizer is adopted to train models in 300 epochs, with an initial learning rate of 0.001.

4. Experiments

In this section, we evaluate our method on two public walking pedestrian video datasets: ETH and UCY. These

Table 1. Comparison with baselines models on ADE & FDE evaluation metrics. † denotes that the scene information is used in this model.

Model	Performance (ADE/FDE) ↓					
	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVERAGE
S-LSTM [2]	1.09/2.35	0.79 / 1.76	0.67 / 1.40	0.47 / 1.00	0.56 / 1.17	0.72 / 1.54
CIDNN [33]	1.25 / 2.32	1.31 / 2.36	0.90 / 1.86	0.50 / 1.04	0.51 / 1.07	0.89 / 1.73
SGAN [10]	0.81 / 1.52	0.72 / 1.61	0.60 / 1.26	0.34 / 0.69	0.42 / 0.84	0.58 / 1.18
SoPhie† [25]	0.70 / 1.43	0.76 / 1.67	0.54 / 1.24	0.30 / 0.63	0.38 / 0.78	0.54 / 1.15
IDL [21]	0.59 / 1.30	0.46 / 0.83	0.51 / 1.27	**0.22 / 0.49**	**0.23 / 0.55**	0.40 / 0.89
STGAT [14]	0.65 / 1.12	0.35 / 0.66	0.52 / 1.10	0.34 / 0.69	0.29 / 0.60	0.43 / 0.83
RAMP† [27]	0.69 / 1.24	0.43 / 0.87	0.53 / 1.17	0.28 / 0.61	0.28 / 0.59	0.44 / 0.90
TPNet† [7]	0.84 / 1.73	0.24 / 0.46	**0.42 / 0.94**	0.33 / 0.75	0.26 / 0.60	0.42 / 0.90
NMMP [13]	0.61 / 1.08	0.33 / 0.63	0.52 / 1.11	0.32 / 0.66	0.29 / 0.61	0.41 / 0.82
STIRNet	**0.48 / 0.95**	**0.22 / 0.41**	0.54 / 1.15	0.37 / 0.80	0.31 / 0.70	**0.38 / 0.80**

two datasets contain 5 crowd scenes, including ETH, HOTEL, ZARA1, ZARA2, and UNIV. There are 1536 pedestrians and thousands of real-world pedestrian trajectories. All the trajectories are converted to the world coordinate system and then interpolated to obtain values at every 0.4 seconds.

Experiment Setup. We use the leave-one-out approach similar to that from S-LSTM [2]. Specifically, we train models on four datasets and test on the remaining dataset. We take the coordinates of 8 key frames (3.2s) of the pedestrian as the observed trajectory, and predict the trajectory of the next 12 key frames (4.8s). For each mini-batch, random rotation is employed for data augmentation.

Evaluation Metrics. Similar to prior works [10, 38], the proposed method is evaluated with two types of metrics as follows:

1. *Average Displacement error(ADE)*: the mean square error (MSE) between the ground-truth trajectory and predicted trajectory over all predicted time steps.

2. *Final Displacement error(FDE)*: the mean square error (MSE) between the ground-truth trajectory and predicted trajectory at the last predicted time steps.

Baselines. We compare the proposed model with the following state-of-the-art models:

1. *S-LSTM* [2]: A recursive trajectory prediction model via LSTM which uses a social pool module to model social interactions.

2. *CIDNN* [33]: A recursive trajectory prediction model which models crowd interactions via spatial affinity.

3. *SGAN* [10]: A GAN-based seq2seq trajectory prediction model that can generate multiple socially acceptable trajectories, in which global pooling is used for social interaction modeling.

4. *SoPhie* [25]: An improved version of SGAN that coupling attention to social and physical constraints.

5. *IDL* [21]: A novel imitative decision learning approach for multimodal path forecasting which delves deeper into the latent decision.

6. *STGAT* [14]: A seq2seq trajectory prediction model which models spatial interaction via GAT and utilizes LSTM for temporal modeling from spatial interaction contexts to capture spatio-temporal interactions.

7. *RAMP* [27]: An improved version of SGAN by coupling extra scene information, in which the forward and backward prediction networks are tightly coupled and satisfying the reciprocal constraint.

8. *TPNet* [7]: A unified two-stage motion prediction framework for both vehicles and pedestrians.

9. *NMMP* [13]: An improved version of SGAN which uses a novel neural motion message passing to explicitly model the interaction and learn representations for directed interactions between actors.

4.1. Quantitative Evaluation

We compare our method with the state-of-the-art baselines mentioned above. All the stochastic method samples 20 times and reports the best-performed sample. The main results are presented in Table 1. The S-LSTM, CIDNN, and the proposed STIRNet are recursive structured models while the rest of baselines are seq2seq models. The performance of STIRNet model is best on ETH and HOTEL datasets and compatible on the rest 3 datesets. STIRNet improves the state-of-the-art prediction to 0.38m and 0.80m on ADE and FDE on average. Particularly, the SoPhie, RAMP, and TPNet models adopt scene information in modeling, but our model achieves better performance without using scene information compared with these models.

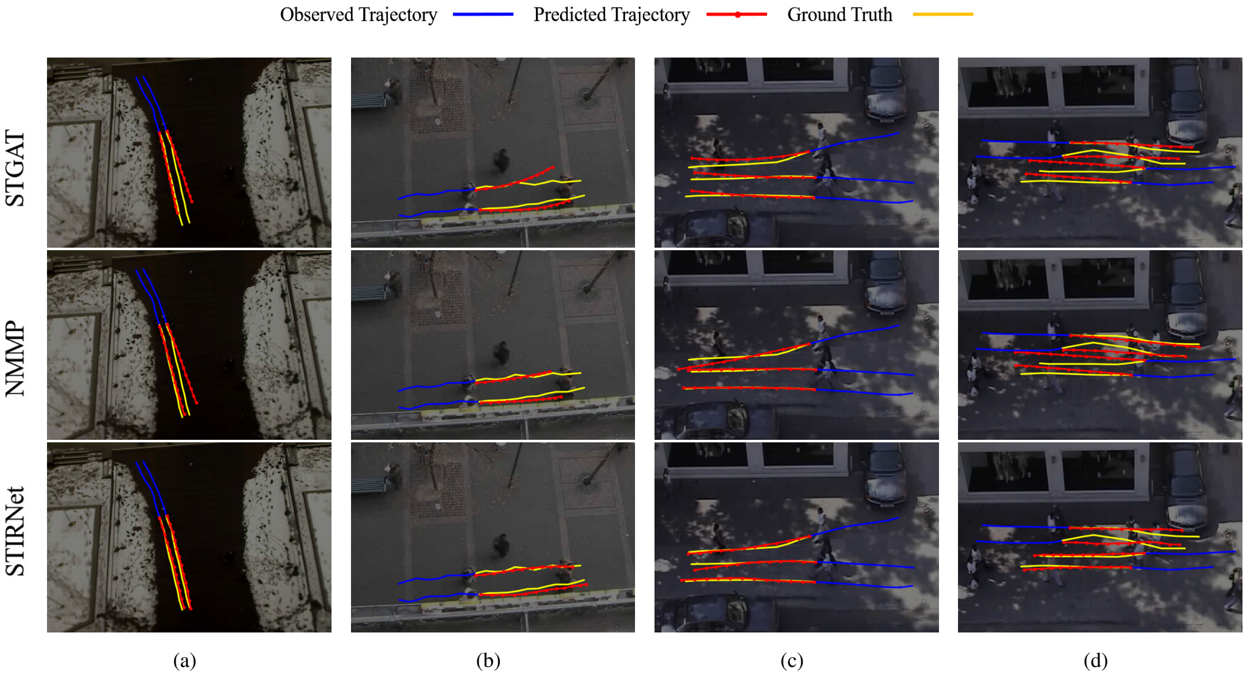


Figure 3. Comparisons between our model with STGAT and NMMP models in 4 different scenarios, which contain parallel walking (a, b), people merging (c), and people meeting (d). For a better view, only part of the pedestrians in the scene are presented. We can see that the trajectories generated by our model are closer to groundtruth.

Inference speed and model size To evaluate the inference speed, we list out the size of parameters and inference speed comparisons between our model and publicly available models which we could bench-mark against. For evaluating the inference speed, we treat each fragment with 20 time-step as a batch and calculate the average time over all batches. In particular, the parameter size of the NMMP model is different across the five sub-datasets, the value listed in the table is the average value. The parameter size of our model is 49.1K, which is about 1.06 times that of SGAN, only second to the 115.8K parameters of NMMP. However, the inference speed of STIRNet is the slowest, which is 11.55ms/batch. The reason is that the recursive structure based STIRNet model performs spatio-temporal interaction modeling in every time-step, while other models based on seq2seq structure only model interaction in the encoder or decoder stage.

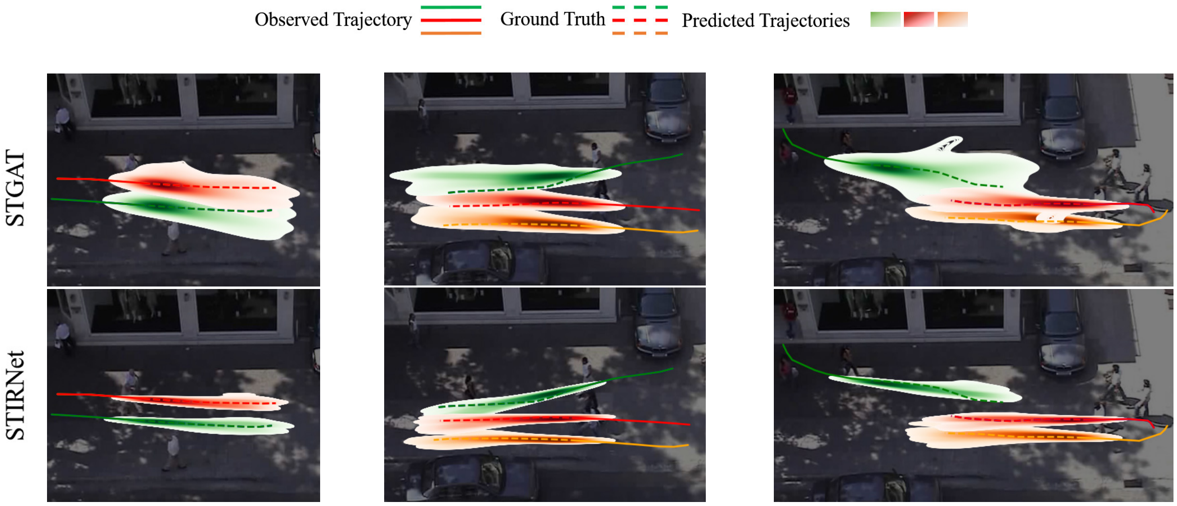
Table 2. Comparisions of parameter amount and inference speed on ETH & UCY datasets. All models evaluated on Nvidia GTX2080Ti GPU.

Model	Parameters (k)	Speed (ms/batch)
SGAN [10]	46.4 (1x)	1.25 (1x)
STGAT [14]	44.6 (0.96x)	1.33 (1.06x)
NMMP [13]	115.8 (2.50x)	4.49 (3.60x)
STIRNet	49.1 (1.06x)	11.55 (9.24x)

4.2. Qualitative Evaluation

As mentioned above, pedestrian trajectory prediction is a complex problem because of the complex social interactions between pedestrians. To verify the effectiveness of our model, we illustrate the prediction trajectories of 4 examples which come from three types of social scenario. The Fig. 3(a) and 3(b) show the parallel walking scenario where two pedestrians are walking in parallel. The trajectories generated by our model are closer to the ground truth while the trajectories predicted by STGAT and NMMP are deviated and fail to reach the endpoints. In people merging Fig. 3(c) and people meeting Fig.3(d) scenarios, the trajectories predicted by our method are also closer to the ground truth and without collisions and crowding happening. These examples prove that the proposed spatio-temporal interaction modeling is more effective and successful than that of the STGAT model.

We also compare the proposed model with STGAT in 3 common social scenarios on multimodal prediction performance (see Figure 4). For the multimodal predictions of the STIRNet model, the ground truth trajectories are always distributed in the high density regions (deep color). Compared with the multimodal prediction of STGAT, the multiple trajectories generated by STIRNet are more concentrated and clustered. However, a wider distribution of future predictions means that there is more randomness in



(a) (b) (c)

Figure 4. Comparisons between our model with STGAT on multimodal predictions. A variety of scenarios are shown: two individuals walking in parallel (a), two persons merging from the same direction (b), and two persons meeting from different directions (c). For each case, the solid lines are observed trajectories, the dashed lines are groundtruth, and the color densities are the predicted trajectory distributions.

the prediction, which is not what we want. Therefore, the prediction distribution generated by STIRNet is more concentrated, which is more efficient.

4.3. Ablation Study

In this section, we verify the effectiveness of our model through ablation studies. The STIRNet model performs spatial modeling and temporal modeling via GAT and LSTM alternately. To verify the validity of this alternate recursive structure, we compare the proposed model with two variants. The TRNet represents a simple version of STIRNet without the GAT-based spatial modeling module. Moreover, a variant model named SIRNet that the temporal nodes features are replaced by spatial features which are acquired from the encoders. As the comparisons listed in Table 3, STIRNet achieves the best performance on ETH, HOTEL, and ZARA2 datasets, and the SIRNet performs best on the rest datasets. Furthermore, STIRNet achieves the best performance on average of five datasets.

We also verify the effectiveness of latent variable generator of STIRNet model in this section. In the training stage, the latent variables of STIRNet model are generated through a VAE model from the concatenated feature $e_i^t \oplus \vec{H}_i^t$ in Sect. 3.4, and the results are listed in column 3 in Table 4. The comparison of a variant version of STIRNet which without VAE model is listed in column 1, while the latent variables are generated from the standard normal distribution. The model via VAE latent variable generator achieves better performance on 4 subdatasets. The results show that using VAE models to generate latent variables is more effective for the training of multimodal prediction models.

Table 3. Ablation study on the effectiveness of alternate recursive structure. SRNet means that the model without the GAT module. SIRNet means that the nodes features in GAT are represented by spatial features.

Dataset	Models		
	TRNet	SIRNet	STIRNet
ETH	0.53 / 1.05	0.50 / 0.99	**0.48 / 0.95**
HOTEL	0.27 / 0.58	0.25 / 0.47	**0.22 / 0.42**
UNIV	0.56 / 1.17	**0.53 / 1.14**	0.54 / 1.15
ZARA1	**0.37** / 0.80	**0.37 / 0.79**	**0.37** / 0.80
ZARA2	0.32 / 0.71	0.34 / 0.73	**0.31 / 0.70**
AVERAGE	0.41 / 0.86	0.40 / 0.82	**0.38 / 0.80**

Table 4. Ablation study of latent variables generator. $\mathcal{N}(0,1)$ means that the latent variables are generated from the standard normal distribution in training stage. $\mathcal{N}(\mu, diag(\delta^2))$ means that the latent variables are generated from a VAE model which are adopted in STIRNet.

Dataset	Latent Variable Generator	
	$\mathcal{N}(0,1)$	$\mathcal{N}(\mu, diag(\delta^2))$
ETH	0.50 / 1.01	**0.48 / 0.95**
HOTEL	0.23 / 0.44	**0.22 / 0.41**
UNIV	0.54 / 1.16	**0.54 / 1.15**
ZARA1	**0.35 / 0.75**	0.37 / 0.80
ZARA2	0.35 / 0.76	**0.31 / 0.70**
AVERAGE	0.39/0.82	**0.38/0.80**

5. Conclusion

In this work we focus on modeling spatio-temporal interaction and jointly predicting trajectories for all people in a scene. We propose a novel spatio-temporal interaction-aware recursive network to predict multimodal socially acceptable trajectories. The ablation studies prove the validity of the proposed spatio-temporal modeling with alternative recursive manner in pedestrian trajectory prediction. The quantitative and qualitative comparisons also verify the effectiveness of the proposed model and outperforms other SOTA methods. Although the proposed STIRNet achieves the state-of-the-art prediction, the inference speed is far less than other models. In future work, we will transfer the proposed spatio-temporal interaction modeling to seq2seq structured model to improve the inference speed.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (61972128), the Fundamental Research Funds for the Central Universities of China (Grant No.PA2019GDPK0071).

References

[1] Priyanshu Agarwal, Suren Kumar, Julian Ryde, and et. al. *Feature-Based Prediction of Trajectories for Socially Compliant Navigation*, pages 193–200. MIT Press, 2013. 2

[2] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, 2016. 1, 2, 4, 5

[3] Javad Amirian, Jean-Bernard Hayer, and Julien Pettre. Social ways: learning multi-modal distributions of pedestrian trajectories with gans. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2964–2972, 2019. 1, 2, 4

[4] Arindam Biswas and Morris Brendan Tran. Tagcn: Topology-aware graph convolutional network for trajectory. In *2020 International Symposium on Visual Computing (ISVC)*, volume 2, pages 542–553, 2020. 2

[5] Patrick Dendorfer, , Aljoša Ošep, and Laura Leal-Taixé. Goal-gan: Multimodal trajectory prediction based on goal position estimation. In *2020 Asian Conference on Computer Vision (ACCV)*, volume 13623, pages 405–420, 2020. 2

[6] David Ellis, Eric Sommerlade, and Ian Reid. Modelling pedestrian trajectory patterns with gaussian processes. In *2009 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1229–1234, 2009. 2

[7] Liangji Fang, Qinhong Jiang, Jianping Shi, and Bolei Zhou. Tpnet: Trajectory proposal network for motion prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6796–6805, 2020. 4, 5

[8] 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. 1

[9] Ascensión Gallardo-Antolín and Juan M. Montero. On combining acoustic and modulation spectrograms in an attention lstm-based system for speech intelligibility level classification. *Neurocomputing*, 456:49–60, 2021. 4

[10] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2255–2264, 2018. 1, 2, 4, 5, 6

[11] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Phys Rev E*, 51(5):4282–4286, 1998. 2

[12] Jiaojiao Hu, Xiaofeng Wang, Ying Zhang, Depeng Zhang, Meng Zhang, and Jianru Xue. Time series prediction method based on variant lstm recurrent neural network. *Neural Processing Letters*, 52:1485–1500, 2020. 4

[13] Yue Hu, Siheng Chen, Ya Zhang, and Xiao Gu. Collaborative motion prediction via neural motion message passing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6318–6327, 2020. 4, 5, 6

[14] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6271–6280, 2019. 1, 2, 3, 4, 5, 6

[15] E. Angus John. Forecasting, structural time series and the kalman filter. *Technometrics*, 34(4):496–497, 1992. 2

[16] Kapil D. Katyal, Gregory D. Hager, and Chien-Ming Huang. Intent-aware pedestrian prediction for adaptive crowd navigation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3277–3283, 2020. 2

[17] Kris M. Kitani, Brian D. Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In *2012 Eupean Conference on Computer Vision (ECCV)*, pages 201–214, 2012. 2

[18] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In *2019 33rd International Conference on Neural Information Processing Systems (NeurIPS)*, pages 137–146, 2019. 1, 2, 3

[19] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B. Choy, Philip H. S. Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2165–2174, 2017. 4

[20] Dongyang Li, Ruimin Hu, Wenxin Huang, Dengshi Li, Xiaochen Wang, and Chenhao Hu. Trajectory association for person re-identification. *Neural Processing Letters*, 2021. 1

[21] Yuke Li. Which way are you going? imitative decision learning for path forecasting in dynamic scenes. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 294–303, 2019. 4, 5

[22] Yuanman Li, Rongqin Liang, Wei Wei, Wei Wang, Jiantao Zhou, and Xia Li. Temporal pyramid network with spatial-temporal attention for pedestrian trajectory prediction. *IEEE*

Transactions on Network Science and Engineering, pages 1–14, 2021. 1, 2

[23] Matthias Luber, Johannes A. Stork, Gian Diego Tipaldi, and Kai O. Arras. People tracking with human motion predictions from social forces. In *2010 IEEE International Conference on Robotics and Automation (ICRA)*, pages 464–469, 2010. 1

[24] Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It is not the journey but the destination : Endpoint conditioned trajectory prediction. In *2020 Eupean Conference on Computer Vision (ECCV)*, volume 1, pages 759–776, 2020. 2

[25] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, S. Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths pompliant to social and physical constraints. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1349–1358, 2019. 2, 5

[26] Yue Song, Niccoló Bisagno, Syed Zohaib Hassan, and Nicola Conci. Ag-gan: An attentive group-aware gan for pedestrian trajectory prediction. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8703–8710, 2021. 2

[27] Hao Sun, Zhiqun Zhao, and Zhihai He. Reciprocal learning networks for human trajectory prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7414–7423, 2020. 2, 5

[28] Jianhua Sun, Qinhong Jiang, and Cewu Lu. Recursive social behavior graph for trajectory prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 657–666, 2020. 2

[29] Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4601–4607, 2018. 2

[30] Chengxin Wang, Shaofeng Cai, and Gary Tan. Graphtcn: Spatio-temporal interaction modeling for human trajectory prediction. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3449–3458, 2021. 2

[31] Ruiping Wang, Yong Cui, Xiao Song, Kai Chen, and Hong Fang. Multi-information-based convolutional neural network with attention mechanism for pedestrian trajectory prediction. *Image and Vision Computing*, 107:104110, 2021. 1

[32] Dan Xiong. Spatial-temporal block and lstm network for pedestrian trajectories prediction. *arXiv:2009.10468*, 2020. 2

[33] Yanyu Xu, Zhixin Piao, and Shenghua Gao. Encoding crowd interaction with deep neural network for pedestrian trajectory prediction. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5275–5284, 2018. 5

[34] Yi Xu, Dongchun Ren, Mingxia Li, Yuehai Chen, Mingyu Fan, and Huaxia Xia. Tra2tra: Trajectory-to-trajectory prediction with a global social spatial-temporal attentive neural network. *IEEE Robotics and Automation Letters*, 6(2):1574–1581, 2021. 1, 2

[35] Yi Xu, Jing Yang, and Shaoyi Du. Cf-lstm: Cascaded feature-based long short-term networks for predicting pedestrian trajectory. In *The 34th AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 12541–12548, 2020. 2, 4

[36] Biao Yang, Guocheng Yan, Pin Wang, Ching-Yao Chan, Xiang Song, and Yang Chen. A novel graph-based trajectory predictor with pseudo-oracle. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2021. 2

[37] Yu Yao, Ella Atkins, Matthew Johnson-roberson, Ram Vasudevan, and Xiaoxiao Du. Bitrap: Bi-directional pedestrian trajectory prediction with multi-modal goal estimation. *IEEE Robotics and Automation Letters*, 6(2):1463–1470, 2021. 2

[38] Pu Zhang, Wangli Ouyang, Pengfei Zhang, Xue Jianru, and Nanning Zheng. Sr-lstm: state refinement for lstm towards pedestrian trajectory prediction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12077–12086, 2019. 2, 4, 5

[39] Hao Zhou, Dongchun Ren, Huaxia Xia, Mingyu Fan, Xu Yang, and Hai Huang. Ast-gnn: An attention-based spatio-temporal graph neural network for interaction-aware pedestrian trajectory prediction. *Neurocomputing*, 445:298–308, 2021. 1, 2, 3