CoMoGCN: Coherent Motion Aware Trajectory Prediction with Graph Representation

Yuying Chen*
ychenco@connect.ust.hk
Congcong Liu*
cliubh@connect.ust.hk
Bertram E. Shi
eebert@ust.hk
Ming Liu
eelium@ust.hk

Robotics Institute
Hong Kong University of Science and Technology
Hong Kong, China

Abstract

Forecasting human trajectories is critical for tasks such as robot crowd navigation and autonomous driving. Modeling social interactions is of great importance for accurate group-wise motion prediction. However, most existing methods do not consider information about coherence within the crowd, but rather only pairwise interactions. In this work, we propose a novel framework, coherent motion aware graph convolutional network (CoMoGCN), for trajectory prediction in crowded scenes with group constraints. First, we cluster pedestrian trajectories into groups according to motion coherence. Then, we use graph convolutional networks to aggregate crowd information efficiently. The CoMoGCN also takes advantage of variational inference to capture the variability in human trajectories by modeling the distribution. Our method achieves state-of-the-art performance on several different trajectory prediction benchmarks, and the best average performance among all benchmarks considered.

1 Introduction

Forecasting human trajectories is of great importance for tasks, such as robot navigation in crowds, autonomous driving, and crowd surveillance. For autonomous robot systems, predicting the human motion enables feasible and efficient planning and control.

However, making accurate trajectory predictions is still a challenging task because pedestrian trajectories can be affected by many factors, such as the topology of the environment, intended goals, and *social relationships and interactions* [20]. Furthermore, the highly *dynamic* and *multimodal* properties inherent in human motion must also be considered.

Multimodality in trajectory prediction has been studied recently [2, 7, 13, 14, 21]. Most past work uses generative adversarial models (GANs) to generate multiple predictions. However, GANs suffer from the instability of adversarial training, which is sensitive to hyperparameters and structure [26]. As an alternative, variational autoencoder (VAE) is relatively

* Equal Contribution

more stable. Lee *et al*. present a CVAE based framework to predict future object locations [14]. A recent work adopted CVAE for trajectory prediction [11]. This paper takes advantage of a VAE-like model to capture the variability of human trajectories.

Recently, some works have proposed to model the dynamic interactions of pedestrians by combining information from pairwise interactions, through pooling mechanisms such as max-pooling [7] and self-attention pooling [21]. However, those works do not completely capture important information about the geometric configuration of the crowd. Furthermore, these works rely on ad-hoc rules to handle varying numbers of agents, such as setting a maximum on the number of agents and using dummy values for non-existing agents [21]. To avoid such ad-hoc assumptions, Chen *et al*. [5] proposed to use graph convolutional networks (GCN) to aggregate information about neighboring humans for robot crowd navigation tasks. The GCN can handle varying numbers of neighbors naturally, and can modulate interactions by changing its adjacency matrix. In this paper, we use a similar graph structure for crowd information aggregation in a different task: trajectory prediction.

Most previous work has focused only on the interactions between pairs of humans. Coherent motion patterns of pedestrian groups, which encode rich information about implicit social rules, has rarely been considered. This lack of attention may be due in part to the lack of information about social grouping in current benchmark datasets, such as the commonly used **ETH [19] and UCY [15]** datasets, for trajectory prediction. To address this unavailability, we add coherent motion cluster labels to trajectory prediction datasets using a coherent filtering method [29], and leverage DBSCAN clustering to compensate for the drawbacks of the coherent filtering method in the small group detection. These coherent motion labels provide a mid-level representation of crowd dynamics, which is very useful for crowd analysis. We incorporated the coherent motion constraints into our model by using GCNs for intergroup and intragroup relationship modeling.

There are several main contributions of our work:

- Unlike past work that considered pairwise interactions between individuals only, we take into account coherent motion constraints inside crowds to better capture social interactions.
- We developed a hybrid labeling method to add coherent motion labels to trajectory prediction datasets. We have released the re-labelled dataset publicly for use by other researchers[1].
- Incorporating the coherent motion into GCN for interaction modeling, the CoMoGCN achieves state-of-the-art performance on several different trajectory prediction benchmarks, and the best average performance across all datasets considered.

2 Related works

2.1 Crowd Interaction

A pioneering work for crowd interaction modeling, the Social Force Model (SFM) proposed by [9], has been applied successfully to many applications such as abnormal crowd behavior detection [17] and multi-object tracking [19]. However, as discussed in [1], the social force model can model simple interactions, but fails to model complex crowd interactions. There

[1] https://comogcn.ram-lab.com

are also other hand crafted feature based models, such as continuum dynamics [23], discrete choice [3] and Gaussian Process models [24]. However, all the above methods are based on hand-crafted energy functions and specific rules, which limit their performance.

2.2 RNN for Trajectory Prediction

Recently, Recurrent Neural Networks (RNN), such as the Long Short Term Memory (LSTM), have achieved many successes in trajectory prediction tasks [1, 8, 16, 22, 27, 28]. Alahi *et al*. proposed a social pooling layer to model neighboring humans [1]. Gupta *et al*. proposed a pooling module, which consists of an MLP followed by max-pooling to aggregate information from all other humans [7]. Sadeghian *et al*. [21] adopted a soft attention module to aggregate information across agents. More recent work uses GCNs to aggregate information by treating humans as nodes and modeling interaction through edge strength for robot navigation [5]. Similarly, a variant of the GCN, the Graph Attention Network (GAT), has been used to model the social interactions [10, 13]. However, the use of multi-head attention in the GAT increases the number of parameters and the computational complexity of the GAT in comparison to the GCN. In this work, we integrate information across humans using GCNs, which enables our method to handle varying crowd sizes.

2.3 Coherent Motion Information for Motion Prediction

Most previous work only pay attention to interactions among pairs of pedestrians. However, the pedestrian trajectories are also influenced by more complex group-wise social relations. Coherent motion patterns inside crowds, which encode implicit social information, have been shown to be useful in many applications, such as crowd activity recognition[25]. Bisagno *et al*. [2] considered intragroup interactions for trajectory predictions, but neglected intergroup interactions. Current benchmark datasets for trajectory prediction do not provide coherent motion labels.

Several works have been addressed detecting coherent motion [29] and measuring the collectiveness of crowds [18]. Zhou *et al*. [29] proposed coherent filtering, which detects invariant neighbors of every individual, and measures the velocity correlations, for motion clustering. It shows good performance on a collective motion benchmark and can detect coherent motions given the crowd trajectories in a short time window. In this paper, we use the coherent filtering method to label trajectory prediction datasets. In addition, we leverage DBSCAN clustering to compensate for the disadvantages of the coherent filtering method in small group detection. Based on the labels, we incorporate coherent motion information into our model for better interaction modeling.

3 Method

3.1 Problem Definition

The goal of this work is to generate the future trajectories of all humans in a scene at the same time. The trajectory of a person i is defined using $x^t_{\text{rel}_i} = (x^t_i, y^t_i)$ which denotes the relative position of human i at time step t to the position at $t-1$. Consistent with previous works [7, 21], the observed trajectory of all humans in a scene is defined as $x^{(1:t_{\text{obs}})}_{\text{rel}_{1,\ldots,N}}$ for time steps $t = 1,\ldots,t_{\text{obs}}$; the future trajectory to be predicted is defined as $x^{(t_{\text{obs}}+1:t_{\text{obs}}+T)}_{\text{rel}_{1,\ldots,N}}$ for time

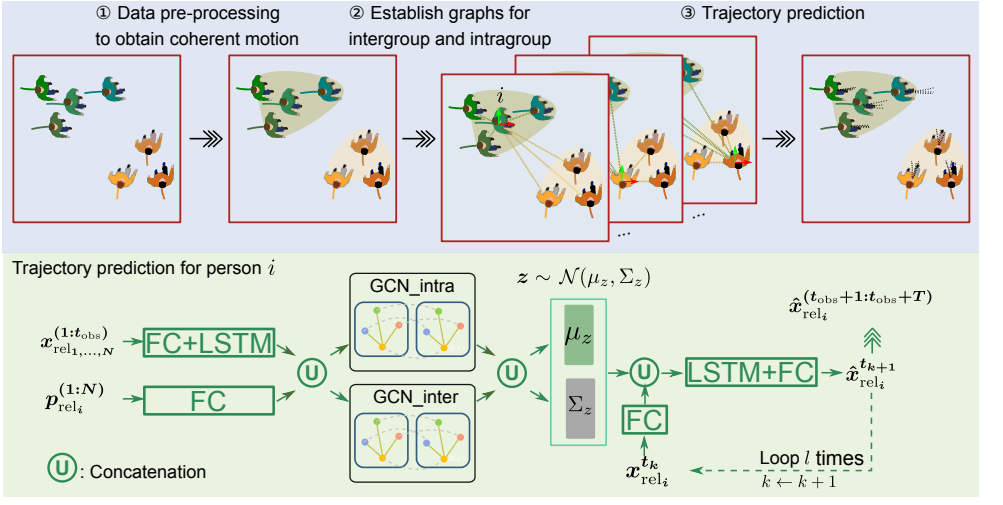

Figure 1: System overview. There are three procedures: 1. We obtain coherent motion labels for each human in an offline data pre-processing procedure. 2. Based on the coherent motion labels for each human, we establish graphs capturing intergroup and intragroup relationships. 3. The encoder LSTM takes past trajectories as input and feeds the encoded features into two GCNs. The embeddings from the two GCNs are concatenated and forwarded to an MLP to create a distribution with mean μ_z and variance Σ_z. Then, features are sampled from the distribution and fed into a decoder LSTM for trajectory prediction.

step $t = t_{\mathrm{obs}}+1, \ldots, t_{\mathrm{obs}}+T$, where the number of humans N may change dynamically. The model aims to generate trajectories $\hat{x}_{\mathrm{rel}_{1,\ldots,N}}^{(t_{\mathrm{obs}}+1:t_{\mathrm{obs}}+T)}$ whose distribution matches that of ground truth future trajectories of all humans $x_{\mathrm{rel}_{1,\ldots,N}}^{(t_{\mathrm{obs}}+1:t_{\mathrm{obs}}+T)}$.

3.2 Overall Framework

The upper half of Fig. 1 shows the overall framework of our method for trajectory prediction. Data pre-processing is applied offline to obtain the coherent motion pattern of the pedestrians. Based on the labels of coherent motion clusters, we establish intergroup and intragroup graphs for each pedestrian. These graphs are utilized in the following trajectory prediction as an efficient way of information aggregation.

3.3 Trajectory Prediction

The lower half of Fig. 1 shows our trajectory prediction model. For simplicity, we show the prediction process for person i's trajectory. The prediction process for others is similar.

For feature extraction, we first use a single layer MLP (FC) to encode each pedestrian's relative displacements as a fixed-length embedding. These embeddings are fed to an LSTM as shown below:

$$e_i = LSTM_{\mathrm{en}}(MLP_{\mathrm{enc}}(x_{\mathrm{rel}_i}; W_{\mathrm{enc}}), h_{\mathrm{enc}_i}, W_{\mathrm{en}}) \tag{1}$$

where W_{enc} contains the weights of the FC layer, and W_{en} contains the weights of the encoding LSTM. The positions of humans $1, \ldots, N$ relative to person i, $p_{\mathrm{rel}_i}^{(1:N)}$, are calculated given

the coordinates of all the pedestrians at t_{obs} in the world coordinate system. These are fed into an FC layer which is similar to the pooling module in Social GAN [7], to obtain the social information p_i.

The features from all pedestrians $e_{1,\ldots,N}$ and the social information for person i, p_i, are concatenated together as the input to two GCNs: one for intragroup and one for intergroup interaction aggregation:

$$V_{\text{intra}_i} = GCN_{\text{intra}}([e_{1,\ldots,N}, p_i], A_{\text{intra}}, W_{\text{intra}}) \tag{2}$$

$$V_{\text{inter}_i} = GCN_{\text{inter}}([e_{1,\ldots,N}, p_i], A_{\text{inter}}, W_{\text{inter}}) \tag{3}$$

where A_{intra} and A_{inter} denote the adjacency matrices as described in more detail in Section 3.5. W_{intra} and W_{inter} are weight matrices. We extract the features of node i after the final graph convolutional layer as the features V_{intra_i} and V_{inter_i}.

The features computed by the outputs of the two GCNs are then concatenated together and input to an MLP, which computes the mean and variance of a distribution over the feature vectors to be input to the decoder:

$$\mu_z, \Sigma_z = MLP_{\text{va}}([V_{\text{intra}_i}, V_{\text{inter}_i}], W_{\text{va}}) \tag{4}$$

where W_{va} is the weight matrix. We sample an input feature vector to the decoder stage, z, from this distribution $z \sim \mathcal{N}(\mu_z, \Sigma_z)$ and concatenate it with the embedding computed from an embedding of the last predicted state. The resulting features c are fed into the decoder LSTM cell for trajectory prediction:

$$\hat{x}_{\text{rel}_i} = MLP_{\text{dec}}(LSTM_{\text{de}}(c, h_{\text{de}_i}; W_{\text{de}}); W_{\text{dec}}) \tag{5}$$

where W_{de} is the weight matrix for the decoder LSTM and W_{dec} is the weight matrix for the decoder MLP.

We minimize the loss for the trajectory prediction:

$$L_{\text{pred}} = ||x_{\text{rel}_i} - \hat{x}_{\text{rel}_i}||_2^2 + \alpha KL(z, q). \tag{6}$$

where $q \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ is a prior distribution and α is a weighting factor. The KL loss forces the distribution of the mean and covariance matrix of (4) close to a normal distribution.

3.4 Coherent Motion Clustering for Pedestrian Groups

For coherent motion detection, we use the coherent filtering proposed by [29]. The process takes the positions of humans from consecutive frames t_1 to t_k and generates a clustering index for each human and for each frame. Humans sharing the same index are considered to have coherent motion. The process of coherent filtering mainly includes three steps: a) finding K nearest neighbors b) finding the invariant neighbors of an individual c) measuring the average velocity correlations over time between the invariant neighbors and the individual. Individual-neighbor pairs with correlation intensity above a threshold are marked as coherent pairs.

Although this method is effective for crowds with large crowd densities, it performs poorly for sparse crowds and fails to detect small groups (as illustrated in Fig. 2(a)). To compensate, we apply an extra clustering step, the DBSCAN method [6], to the unlabeled humans. As a density based clustering method, it relies on the distance to find the neighbors.

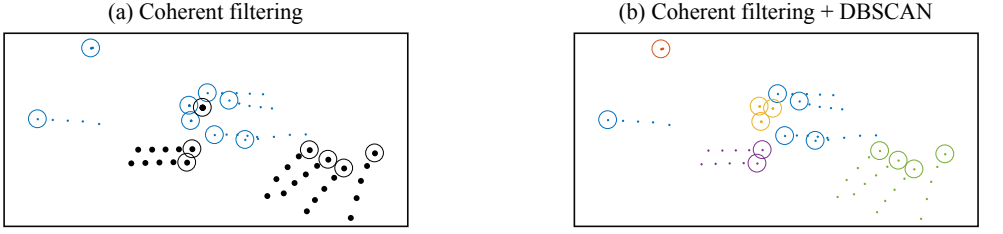

Figure 2: A representative example of coherent motion clustering. Circles show the current position. Dots show the trajectory history used for clustering. Circles with the same color belong to the same group. Black circles have no detected coherence. Compared with the coherent filtering, the hybrid labeling method detects small groups (shown in purple and green) and corrects incorrect labels of static pedestrians (shown in yellow and red). Best viewed in color.

We account for moving direction and calculate the angular distance of each pair of humans. These differences are used to classify humans into clusters.

As shown in Fig. 2, our hybrid labeling method improves the labeling yield and generates better labels than the coherent filtering alone. We also evaluated these two labeling methods quantitatively and obtained consistent results. More details are in the supplementary file, where Table 1 lists the parameter settings of the two methods, Table 2 and 3 shows the quantitative results, and Fig. 1 and Fig. 2 give more clustering examples.

3.5 Intragroup and Intergroup Graph Convolutional Networks

Dealing with the large and varied numbers of humans in a scene is one of the main challenges for multi-human trajectory prediction. Previous works adopted ad-hoc solutions such as setting a maximum number of humans [21]. In this work, we address this problem in a simpler and more principled way through graph representations. Nodes in the graph denote humans in the crowd. In the following, we denote the number of humans in the crowd by N.

We adopt a two-layer graph convolutional networks (GCNs) [12] to aggregate information in crowds. To each node in the network, we associate a feature vector. The graph convolutional layer takes input feature vectors for each node and converts them to output feature vectors for each node by integrating information both within and across nodes. We use I to denote the dimension of the input feature vectors and O to denote the dimension of the output feature vectors The input feature vectors of layer l are represented by matrix $H^l \in \mathbb{R}^{N\times I}$. The input feature matrix is converted to output vectors represented by a matrix $H^{l+1} \in \mathbb{R}^{N\times O}$ based on the layer-wise forward rule:

$$H^{l+1} = \sigma\left(AH^lW^l\right) \tag{7}$$

$W^l \in \mathbb{R}^{I\times O}$ is a trainable weight matrix for layer l. $A \in \mathbb{R}^{N\times N}$ is the adjacency matrix of the graph, whose values determine how information from different nodes is aggregated. Each row of A is normalized to sum to one. $\sigma(\cdot)$ is ReLU activation function.

The adjacency matrix reflects the connections between nodes of the graph. A single vanilla GCN assumes that the influence of each human on another (as determined by W^l) is the same. Only the strength of that influence can be modulated (through the adjacency

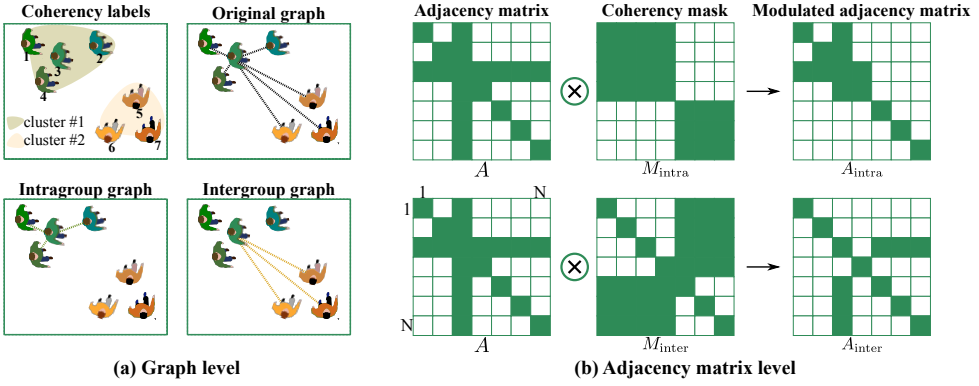

Figure 3: An example of how the adjacency matrices of the GCNs for crowd information aggregation are determined. The example considers the adjacency matrix for the GCNs of human i = 3, who is in the same cluster as humans 1,2 and 4, but not humans 5, 6 and 7. The coherency masks are generated from the coherency labels. A, A_{intra} and A_{inter} correspond to the original graph, intragroup graph and intergroup graph, respectively.

matrix). However, we think that different humans will have different influences on each other, depending on whether the humans are in the same group or not. Thus, we propose to use two GCNs.

As shown in Fig. 3, for each human, we create two adjacency matrices, A_{intra} and A_{inter} using with two coherence masks $M_{\text{intra}}(i,j)$ and $M_{\text{inter}}(i,j)$ which encode the intragroup and intergroup labels (the coherency labels in (a)). Here, i and j are the row and column index, respectively. Each person has his/her own adjacency matrix A, corresponding to a complete star-topology graph with the person at the center (the orginal graph in (a)). If person i and person j share the same (different) label, then $M_{\text{intra}}(i,j)$ will be 1 (0) and $M_{\text{inter}}(i,j)$ will be 0 (1). If i equals j, $M_{\text{intra}}(i,j)$ and $M_{\text{inter}}(i,j)$ equal to 1 to keep the self connection. We obtain the two adjacency matrices by pixelwise multiplying the adjacency matrix (A) with the masks. The resulting A_{intra} and A_{inter} correspond to two star-topology subgraphs (the intragroup and intergroup graph in (a)) with edges selected by the masks. We set the value in the adjacency matrix by first constructing a binary matrix specifying connections between nodes, and then normalizing each row.

By modulating the adjacency matrix of GCNs with coherent motion information, we incorporate implicit social relations into our network for better interaction modeling.

3.6 Implementation Details

We trained the network with Adam optimizer. The mini-batch size is 64 and the learning rate is 1e-4. The models were trained for 200 epochs. MLP_{enc} had a single layer with output dimension 16. The hidden dimension of $LSTM_{\text{en}}$ was 32. The two GCNs had two layers with dimensions 72 and 8. MLP_{va} generated the mean and variance of an 8 dimensional random variable z. $LSTM_{\text{de}}$ had hidden dimension 32. MLP_{dec} had output dimension 2.

4 Experiments

We evaluated our method on two public datasets ETH [19] and UCY [15]. The ETH dataset contain two scenes (ETH and Hotel). The UCY dataset contain three scenes (Zara1, Zara2, and Univ). There are five sets of data with four different scenarios and 1536 pedestrians in total. We follow the S-GAN's data loader and predict trajectories for persons that exist persistently in the observation and prediction time windows.

4.1 Evaluation Methodology

Following the setting in [7], we adopt the leave-one-out approach, i.e. train with four sets and test on the remaining set. We take trajectories of 8 time steps as the observations and evaluate trajectory predictions over the next 12 time steps.

4.1.1 Metrics

Similar to previous works [7, 13, 21], we adopt two standard metrics including Average Displacement Error (ADE) and Final Displacement Error (FDE) in meter. We use the same best-of-N metric as in S-GAN[7].

ADE: Mean L2 distance between the ground truth and the predictions of all time steps.

FDE: Mean L2 distance between the ground truth and the prediction at the final time step.

4.1.2 Baselines

We compare our work with following several recent works based on generative models:

Social GAN (S-GAN) [7]: A generative model using GAN to generate multimodal predictions. It utilizes a global pooling module to combine crowd interactions by an MLP followed by a max-pooling layer.

SoPhie [21]: A improved GAN based model which considers both social interactions and physical interaction with scene context.

Trajectron [11]: A generative model based on CVAE for multimodal predictions with spatiotemporal graphs.

Social-BiGAT [13]: A generative model using Bicycle-GAN for multimodal prediction and GAT for crowd interaction modeling.

4.2 Quantitative results

4.2.1 Comparison to state-of-the-art methods

As shown in Table 1, we compare our models with various baselines. The average displacement error (ADE) and final displacement error (FDE) were reported across five datasets. We ran 20 samples for evaluation.

Our model with GCN and coherent motion constraints outperforms all baselines, with consistently lower values of both ADE and FDE. Compared to Social GAN, we achieve 22.4% improvement in ADE and 22.9% improvement in FDE on average. Compared to SoPhie, which uses additional scene context information, we achieve 16.7% improvement in ADE and 20.9% improvement in FDE on average. Compared to Trajectron, which uses VAE as backbone network, we achieve 15.1% improvement in ADE and 14.2% improvement

Dataset	Baselines				Ours			
	S-GAN	SoPhie	Trajectron	Social-BiGAT	MLP	GCN	GCN+group (CF)	GCN+group (Hybrid)
ETH	0.81/1.52	0.70/1.43	**0.59/1.17**	0.69/1.29	0.73/1.40	0.72/1.31	0.71/1.28	0.70/1.26
HOTEL	0.72/1.61	0.76/1.67	0.42/0.80	0.49/1.01	0.45/0.93	0.41/0.81	0.37/0.76	**0.37/0.75**
UNIV	0.60/1.26	0.54/1.24	0.59/1.21	0.55/1.32	0.61/1.31	0.55/1.18	0.55/1.19	**0.53/1.16**
ZARA1	0.34/0.69	0.30/0.63	0.55/1.09	**0.30/0.62**	0.34/0.72	0.35/0.74	0.34/0.72	0.34/0.71
ZARA2	0.42/0.84	0.38/0.78	0.52/1.04	0.36/0.75	0.33/0.71	0.32/0.68	0.32/0.68	**0.31/0.67**
AVG	0.58/1.18	0.54/1.15	0.53/1.06	0.48/1.00	0.49/1.01	0.47/0.94	0.46/0.93	**0.45/0.91**

Table 1: Quantitative results. We adopted two metrics Average Displacement Error (ADE) and Final Displacement Error (FED) for evaluation over five different datasets (ADE/FDE in meters). Our full model (GCN +group (hybrid)) achieves state-of-the-art results outperforming all baseline methods (lower value denotes better performance).

in FDE on average. Compare to Social-BiGAT, which also considers graph structure for interaction modeling, we achieve 6.3% improvement in ADE and 9.0% improvement in FDE on average.

4.2.2 Ablation study

We conducted several ablation studies to validate the benefits of the use of GCN and coherent motion information. We report average results over multiple runs in Table 1.

To show the benefit of the use of GCN, we investigated a model that replaces the GCN with an MLP (followed by max-pooling, similar to the pooling module in social GAN [7]). The model with GCN achieves 4.1% improvement in ADE and 6.9% improvement in FDE on average.

When comparing the full model with the one using GCN only, the full model with coherent motion information achieves 4.3% improvement in ADE and 3.2% improvement in FDE on average.

The above ablation studies clearly demonstrate the benefits of the use of GCN and the introduction of coherent motion information. We observe consistent improvements over most datasets.

We further investigated trajectory prediction performance of models with different coherent detection methods: the Coherent Filtering method (CF) [29] vs. our hybrid labeling method (hybrid). The model with our hybrid labeling method outperforms the model with only the Coherent Filtering method by 2.2 % in ADE and 2.2 % in FDE on average. The improvements are consistent over all five datasets.

4.3 Qualitative results

To provide better understanding the benefits of our model in capturing social interactions between humans, Fig. 4 shows several examples of the generated trajectories from the testing sets.

From (a) and (b), we can see that in narrow or crowded environments, where interactions is unavoidable, the predictions of our model generally have lower variance than S-GAN. This is expected as the introduction of social interaction models introduces additional constraints, which lower the variance within modes. For those cases where social interactions have little influence (the orange trajectories in (c) and green trajectories in (d)), predictions of our model have large variance, similar to S-GAN. This suggests that our model captures the multi-modal nature of pedestrian trajectories by predicting many possible ways that people could

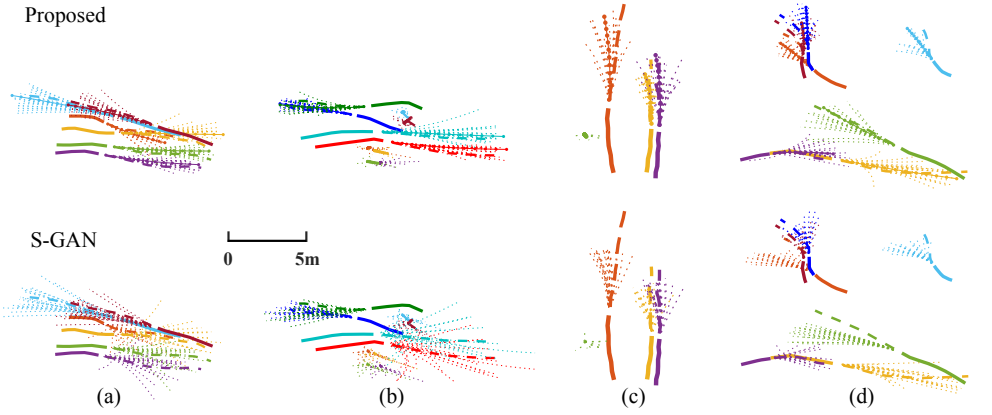

Figure 4: Examples for generated human trajectories visualization for S-GAN and our model across several scenes. The observed trajectories are shown in solid lines, ground truth future trajectories are shown in wide dashed lines, generated 20 samples per model are shown in thin dashed lines. The dot-dashed lines denote the "average" predictions of our model by applying the mean value (μ_z) of the distribution. Different humans are denoted by different colors.

move in the future. Also, the examples show that our model better captures the interactions between pedestrians walking in the crowds which obtain more accurate predictions (as shown in (d)). We also observe that S-GAN tends to predict slower motion in the HOTEL dataset (as shown in (c)).

For qualitative results from the ablation study, please refer to Fig. 3 in supplementary file. We observe results consistent with the quantitative evaluation. Our proposed full model makes more accurate and realistic predictions.

5 Conclusion

In this paper, we proposed a novel VAE-like generative model for trajectory prediction which outperforms state-of-the-art methods. We use graph convolutional networks (GCNs) for efficient crowd interaction aggregation. Furthermore, we have provided coherent motion information for commonly used trajectory prediction datasets (ETH and UCY). These coherent motion labels significantly enrich the social information, and have been released to the research community. Our results show that the introduction of GCNs and coherent motion information significantly improve the performance and accuracy of trajectory prediction.

Acknowledgements This work was supported by the National Natural Science Foundation of China, under grant No. U1713211, Collaborative Research Fund by Research Grants Council Hong Kong, under project No. C4063-18G, the Hong Kong Research Grants Council, under grant 16213617, and HKUST-SJTU Joint Research Collaboration Fund, under project SJTU20EG03.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social LSTM: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–971, 2016.

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

[3] Gianluca Antonini, Michel Bierlaire, and Mats Weber. Discrete choice models of pedestrian walking behavior. *Transportation Research Part B: Methodological*, 40(8): 667–687, 2006.

[4] Niccolo Bisagno, Bo Zhang, and Nicola Conci. Group LSTM: Group trajectory prediction in crowded scenarios. In *The European Conference on Computer Vision Workshops*, September 2018.

[5] Yuying Chen, Congcong Liu, Bertram E Shi, and Ming Liu. Robot navigation in crowds by graph convolutional networks with attention learned from human gaze. *IEEE Robotics and Automation Letters*, 5(2):2754–2761, 2020.

[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, volume 96, pages 226–231, 1996.

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

[8] Irtiza Hasan, Francesco Setti, Theodore Tsesmelis, Alessio Del Bue, Fabio Galasso, and Marco Cristani. MX-LSTM: mixing tracklets and vislets to jointly forecast trajectories and head poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6067–6076, 2018.

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

[10] Yingfan Huang, HuiKun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. STGAT: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6272–6281, 2019.

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

[12] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[13] 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 *Advances in Neural Information Processing Systems*, pages 137–146, 2019.

[14] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. DESIRE: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 336–345, 2017.

[15] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer Graphics Forum*, volume 26, pages 655–664. Wiley Online Library, 2007.

[16] Matteo Lisotto, Pasquale Coscia, and Lamberto Ballan. Social and scene-aware trajectory prediction in crowded spaces. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[17] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 935–942. IEEE, 2009.

[18] Ling Mei, Jianghuang Lai, Zeyu Chen, and Xiaohua Xie. Measuring crowd collectiveness via global motion correlation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

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

[20] Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. Human motion trajectory prediction: A survey. *The International Journal of Robotics Research*, 39(8):895–935, 2020.

[21] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. SoPhie: An attentive GAN for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1349–1358, 2019.

[22] Hang Su, Jun Zhu, Yinpeng Dong, and Bo Zhang. Forecast the plausible paths in crowd scenes. In *International Joint Conferences on Artificial Intelligence*, volume 1, page 2, 2017.

[23] Adrien Treuille, Seth Cooper, and Zoran Popović. Continuum crowds. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 1160–1168. ACM, 2006.

[24] Jack M Wang, David J Fleet, and Aaron Hertzmann. Gaussian process dynamical models for human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):283–298, 2007.

[25] Xiaogang Wang, Xiaoxu Ma, and W Eric L Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(3):539–555, 2008.

[26] Zhengwei Wang, Qi She, and Tomas E Ward. Generative adversarial networks: A survey and taxonomy. *arXiv preprint arXiv:1906.01529*, 2019.

[27] Yanyu Xu, Zhixin Piao, and Shenghua Gao. Encoding crowd interaction with deep neural network for pedestrian trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5275–5284, 2018.

[28] Pu Zhang, Wanli Ouyang, Pengfei Zhang, Jianru Xue, and Nanning Zheng. SR-LSTM: State refinement for LSTM towards pedestrian trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12085–12094, 2019.

[29] Bolei Zhou, Xiaoou Tang, and Xiaogang Wang. Coherent filtering: Detecting coherent motions from crowd clutters. In *European Conference on Computer Vision*, pages 857–871. Springer, 2012.