GSMNet: Towards Long-term Trajectory Prediction by Integrating Multi-Scale Information

Shaohua Liu[1 [0000−0001−6042−6941]], Yisu Wang[1,2], Yinglong Zhu[1,2], Pengfei Yao[2,3], Tianlu Mao[2⋆ [0000−0003−2537−9873]], and Zhaoqi Wang[2]

[1] Beijing University of Posts and Telecommunications, China
liushaohua@bupt.edu.cn, yswang@bupt.cn, z.yinglong@bupt.edu.cn
[2] Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, China
{ltm, zqwang}@ict.ac.cn
[3] Institute of Computing Technology, University of Chinese Academy of Sciences, China
yaopengfei22@mails.ucas.ac.cn

Abstract. Predicting the future trajectories of pedestrians is a vital task for many applications, such as autonomous driving and robot navigation. Most existing methods only predict short-term trajectories. In this paper, we challenge the problem of long-term trajectory prediction. Different from short-term prediction which focus most on the local information, long-term prediction needs to model future trajectory with multi-scale information hierarchically from the multimodal global destination, to mid-distance scene layout limitation, other agent movement and finally the local history motion pattern. The destination reflects pedestrian long-term multimodal goal, the scene layout along with interaction constrains the possible path choice, and history motion pattern guides the future movement. We propose GSMNet, which achieves effective long-term trajectory prediction by integrating multi-scale factors: multimodal goals, scene interaction and motion patterns. We design separate modules to extract different scale features. Multi-layer-perceptron extracts the local-scale feature from history motion pattern. U-Net with attention captures the mid-scale pedestrian-scene correlation feature and goal feature with scene layout at global-scale. Finally, combining multi-scale feature to predict future trajectories. Experiments on SDD dataset and ETH-UCY dataset show that proposed GSMNet outperforms the previous state-of-the-art for both long-term and short-term trajectory prediction task. Qualitative results show GSMNet generates more reasonable trajectories.

Keywords: Long-term Trajectory Prediction · Pedestrian Trajectory Prediction · Multimodal Prediction · Multi-scale Information Combination

⋆ Corresponding author. Email: ltm@ict.ac.cn

1 Introduction

Predicting the future trajectories of different moving agents is an essential task in many applications, such as autonomous driving [3, 6, 18], robot navigation [2, 36], and surveillance system [25, 35]. Currently, most existing methods have a limited prediction time horizon of less than five seconds. Such methods could provide reliable multimodal trajectory predictions in short-term. However, in some applications like autonomous driving, the long-term trajectory prediction is in demand, where longer predictions would affiliate better future path planning and decision making. For example, long-term trajectory prediction can help avoid the path planning module of self-driving vehicles produce sudden maneuvers, such as braking, accelerating, or changing lanes, and reduce the risk of severe accidents. Therefore, there is a practical need for long-term trajectory prediction, which can provide more information and time for future actions.

Predicting the future trajectories of pedestrians requires taking into account various factors that influence their future movements. Short-term future trajectories are mostly influenced by the historical motion pattern and local environment influence like nearby obstacles or agents, long-term trajectories are more influenced by the destination, the global environment information and further motion of agents in the scene. As the forecasting time gets longer, the trajectory becomes more challenging to predict. Especially, the long-term trajectory involves longer moving distance, time and more significant scene changes than the previous short-term trajectory. To effectively take these factors into consideration, the modeling of the future trajectory should be performed in not only the local scale motion patterns, but the middle-distance scale environment interactions and the global scale destinations.

When predicting long-term trajectories, the pedestrian destinations could be distant from observed positions, and a global-scale feature need to be modeled. An accurate prediction of destination can facilitate the trajectory generation process by leading a route to it. Future trajectory to the destination is influenced by the scene layout, which constrains the feasible actions of the pedestrians and affects their decision-making process in the middle scale. For example, pedestrians may change their directions at corners and crossroads, or avoid areas such as woods and buildings. Motion of other agents also influences trajectory, consider the social distance and avoidance of collision. Moreover, the local scale historical motion patterns, such as direction, speed, and acceleration, can help us infer the future motion pattern.

In this work, by combining multi-scale factors including global multimodal **G**oals, mid-range **S**cene interaction and local **M**otion patterns, we propose a long-term trajectory prediction model GSMNet. First, we model the historical trajectory and the scene layout to get the pedestrian-scene spatial correlation features. We apply two U-Net with attention mechanism to extract goal-emphasized correlation features and trajectory-emphasized correlation features separately. Then, we apply the proposed scene constrained goal modeling module which consists of two steps: goal distribution generation step, to generate the goal distribution with the goal-emphasized features and scene layout modification step,

to modify it by the scene layout to obtain multiple goals. After that, we derive the motion pattern feature from the observed history trajectory. Finally, we use the trajectory-emphasized features, motion pattern feature, and the goal to generate the future trajectory. We conducted experiments on ETH-UCY and SDD dataset. Experiment results indicate that our model improves long-term prediction accuracy and it predicts more reasonable goals and trajectories on the scene. GSMNet also performs well for short-term trajectory prediction task. Our main contributions are as follows:

- We propose a trajectory prediction model, named GSMNet, which considers multi-scale influence on future trajectory hierarchically including the multi-modal goals, scene interaction and motion pattern to generate more accurate long-term trajectories.
- We propose scene constrained goal modeling module, using scene layout information to ensure the multimodal goals are in the feasible location and make the predicted long-term trajectories consistent with the environment and social constraints.
- We use two separate U-Net with Attention module to extract pedestrian-scene correlation features. One is used to model the goals for goal prediction and the other one models trajectories for trajectory prediction.

2 Related Work

2.1 Motion-based Trajectory Prediction

Most current trajectory prediction methods model the two-dimension position coordinates of the pedestrians to obtain the pedestrian motion features and the temporal movement correlations. [1] first applied LSTM to model the historical location coordinates to achieve trajectory prediction. Many trajectory prediction methods [9, 10, 12, 30, 41, 44] also model the historical trajectories and take RNN-based networks for temporal modeling. In recent years, methods based on CNN [23, 37], Transformer [34, 42, 43], and MLP [21, 38, 39] have also been applied in pedestrian motion modeling. They also capture the features from the bi-dimensional historical trajectory coordinates and show excellent effects. GSMNet also models the two-dimension historical trajectories to capture the motion pattern of pedestrians.

2.2 Scene-based Trajectory Prediction

The scene information restrains the path planning and the destinations of the pedestrians. There are many methods of scene information modeling. Some works [12, 13, 30, 40] adopted CNN layers to extract visual features from the image of the global scene directly. Some works [5, 15, 28] obtain the scene information from the scene semantic segmentation. Some works [4, 16, 17] construct the statistical heatmap denoting the passing probability of each position and use the statistical heatmap to model the scene impact. The above methods combine the

differently acquired scene features with the pedestrian motion features obtained from the historical trajectory coordinates to get the future trajectory. However, such methods separately model the pedestrian movement and the scene information, neglecting the spatial correlation between the pedestrian positions and the scene layout. [11] proposed the multi-agent tensor fusion module to fuse the motion features and the surrounding scene features. [20] get the scene segmentation and construct the trajectory-on-scene heatmap, which could capture the spatial correlation between pedestrians and the scene.

2.3 Multimodal Trajectory Prediction

Pedestrian movement is flexible, so there may be multiple feasible future paths even if they have the same historical trajectories. To obtain multiple possible ways, some models [9] applied Generative Adversarial Networks (GANs) and proposed variety loss to generate various trajectories. Some methods [11, 12, 30] expand on it, using GANs for multimodal trajectory prediction. Some works [13,24,31,41] applied Conditional Variational Autoencoder (CVAE) and train the network with a Kullback-Leibler divergence (KLD) based loss to obtain multiple future trajectories. Some methods [19, 32] also obtain multimodal trajectories through flow based generative methods. By applying a normalizing flow, multimodal trajectories along with their log-likelyhood of the could be calculated and optimized. Recently, the diffusion models reveals its strong generation ability. [7, 22] applied and optimized the diffusion model to generate multimodal future predictions.

2.4 Long-term Trajectory Prediction

[20] are the first to realize effective long-term trajectory prediction, utilizing a 5-second observed trajectory to predict the subsequent 30-second trajectory. This method proposes the trajectory-on-scene heatmap representation, which captures the relation between the pedestrian and the scene, and they employ epistemic and aleatoric structure for diverse long-term trajectory prediction. Though this method effectively realized long-term trajectory prediction, this method does not consider the modeling of the pedestrian motion patterns, and the pedestrian-scene spatial correlation modeling can be further improved.

3 Methodology

The aim of GSMNet is to predict all pedestrians future trajectories in the scene given the historical trajectories and the scene image. This model predicts multiple goals at first, then predicts the complete trajectories according to the goals. This section describes our GSMNet model, which includes a pedestrian-scene correlation encoder module, motion pattern modeling module, goal modeling module, and trajectory modeling module, as shown in Figure 1.

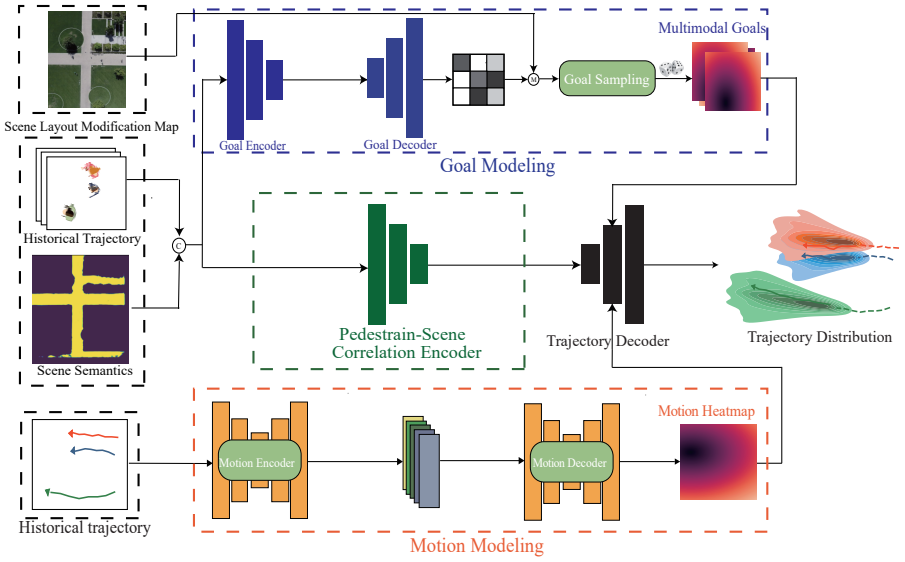

Fig. 1: The architecture of our proposed GSMNet. GSMNet includes a pedestrian-scene correlation encoder to model the relationships between the human and the scene, a motion pattern modeling module to generate future moving tendencies, a goal modeling module to acquire and refine multimodal goals with scene, and a trajectory modeling module to generate the trajectory.

3.1 Problem Definition

Suppose there are N pedestrians in the scene. They are denoted as $p_1, p_2, \ldots, p_N$, the position of pedestrian $p_i (i \in [1, N])$ at timestamp t is denoted as $P_i^t = (x_i^t, y_i^t)$. Given the observed history trajectory $P_i^t (i \in [1, N], t \in [1, T_{obs}])$ and the scene image I, we could acquire the predicted future trajectory $P_i^t (i \in [1, N], t \in [T_{obs+1}, T_{pred}])$, and we aim to make the predicted trajectory close to the future trajectory.

3.2 Pedestrian-scene Correlation Encoder

First, we process the scene image I with a semantic segmentation network to obtain the C class scene semantic segmentation map M_{scene}. Pedestrians may perform different behaviors on different semantic classes. We follow the trajectory-on-scene heatmap in Y-net [20], generate an trajectory heatmap M_{traj} and concatenate it with the scene semantic segmentation map M_{scene} to get the pedestrian-scene heatmap representation M_s. The size of M_{scene} is $H \times W \times C$, the size of M_{traj} is $H \times W \times T_{obs}$, T_{obs} represents the observed frames, and the size of M_s is $H \times W \times (C + T_{obs})$.

$$M_s = Concat(M_{traj}, M_{scene}) \tag{1}$$

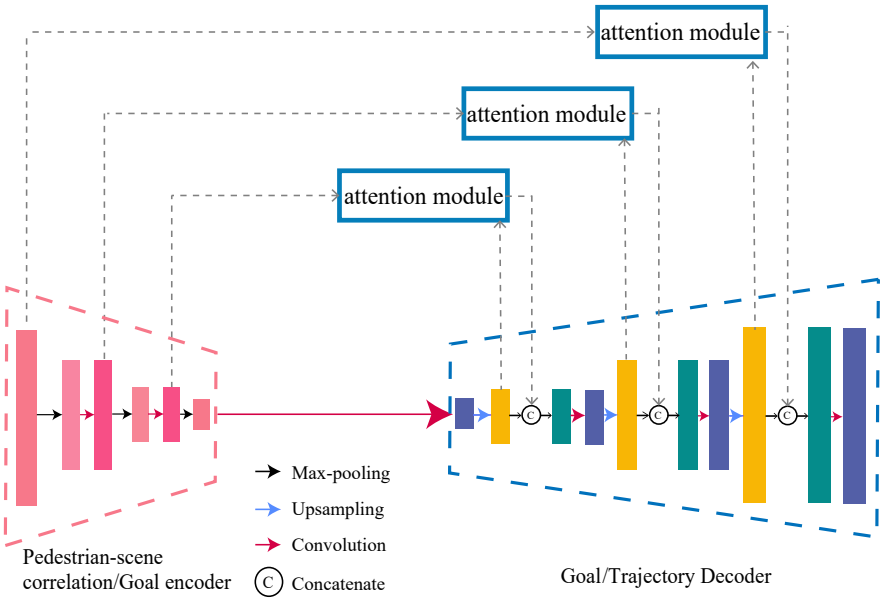

Fig. 2: The encoder-decoder architecture of the pedestrian-scene correlation encoder or the goal encoder and the goal decoder or the trajectory decoder. The two decoders are of similar structure.

We use the pedestrian-scene correlation encoder U_{enc} to obtain the pedestrian-scene correlation feature F.

$$F = U_{enc}(M_s; W_{enc}) \tag{2}$$

Similar to attention U-Net [26], the attention mechanism connects the encoding and decoding convolution layers. The pedestrian-scene correlation encoder encodes the features that are decoded by the goal decoder and the trajectory decoder. These two encoder-decoder structures are the same, as shown in Figure 2. We extract two different pedestrian-scene correlation features for two different tasks: the goal prediction task and the trajectory prediction task. The attention mechanism is used to help obtain those features: F_{des}, which is related to the goal and emphasizes the destination prediction, and F_{traj}, which is related to the trajectory and emphasizes the trajectory prediction.

3.3 Motion Pattern Modeling

The pedestrian-scene heatmap can capture the spatial correlations between pedestrians and the scene, but it is not sufficient to describe the motion patterns of the pedestrians. We model the two-dimensional coordinates of the pedestrian location to learn the motion patterns and habit.

The motion pattern modeling module is shown in the upper part of Figure 1. We encode the two-dimension coordinate sequence of the historical trajectory to obtain the historical motion feature of the pedestrians. After that, we inference the future moving pattern from the history motion feature. We construct the motion encoder to model the motion features, and the pattern decoder to explore the future motion. In Equation 3, f_{enc} denotes the motion encoder, W_{me} denotes the parameters of f_{enc}, f_{dec} is the pattern decoder, W_{md} denotes the parameters of f_{dec}.

$$X_{tend} = f_{dec}(f_{enc}(X_{obs}; W_{me}); W_{md}) \tag{3}$$

3.4 Goal Modeling

Goal distribution generation. The goal generator decodes trajectory-emphasized features encoded by the goal encoder (same structure as pedestrian-scene correlation encoder), and outputs the distribution map of long-term goals. The goal decoder is shown in the right part of Figure 2.

$$M_{des} = U_{des}(F_{des}; W_{des}) \tag{4}$$

Scene Layout Modification. To avoid the goal to appears in unreachable area, we propose the scene layout modification map to facilitate the multimodal goals locates in the reasonable regions in the scene. We analyze the scene layout possibility and manually divide the scene into three categories of pedestrian-scene interaction, namely frequently pass, occasionally pass, and impassable. Frequently pass refers to where pedestrians usually walk on, such as roads or paths. Occasionally pass represents other areas where a few people may choose to walk on, such as lawns. Impassable means buildings, woods, and other places where pedestrians cannot cross.

We manually construct the scene layout modification map based on the three categories above and assign different probabilities to different map categories. We filter out the unachievable goal distribution through the modification map, as shown in Equation 5. M_{sample} is the acquired passable goal distribution, M_{des} is the original goal distribution, S denotes the scene layout modification map containing the accessing probability, $\circ$ representative element-wise multiplication. Finally, sampling is implemented with the sampling method of Y-net [20] to get the finally extracted goal position coordinates X_{goal}.

$$M_{sample} = M_{des} \circ S \tag{5}$$

3.5 Trajectory Modeling

The goal is of vital importance when predicting the complete trajectory. As the destination of pedestrian movement, the goals influence the pedestrians to choose appropriate routes to it. In this method, we integrate the pedestrian's motion pattern, goal, and trajectory-emphasized feature to generate the complete trajectory. Figure 2 represents the pedestrian-scene encoder and trajectory

decoder. Comparing with the goal decoder, the trajectory decoder also combines the motion pattern and goal representation. Therefore, each layer of the trajectory decoder structure has additional two dimensions than the goal decoder. The trajectory decoder outputs the future trajectory distribution map M_{traj}, where the dimension of M_{traj} is $H \times W \times T_{pred}$. T_{pred} represents the predicting period of the future trajectory. Then, softargmax operation is performed on the trajectory distribution map to obtain the corresponding trajectory coordinate X_{traj}, where the dimension of X_{traj} is $2 \times T_{pred}$. In Equation 6 and 7, F_{traj} represents the trajectory-emphasized features, M_{tend} is the pattern heatmap, M_{goal} is the goal heatmap, U_{traj} is the trajectory decoder, M_{traj} is the trajectory distribution map, W_{traj} is the parameter of the trajectory decoder, X_{traj} is the pedestrian trajectory obtained by the statistics of trajectory distribution map.

$$M_{traj} = U_{traj}(Concat(F_{traj}, M_{tend}, M_{goal}); W_{traj}) \tag{6}$$

$$X_{traj} = softargmax(M_{traj}) \tag{7}$$

4 Experiments and Results

Datasets We conduct experiments on two real-life datasets, Stanford Drone Dataset (SDD) [29] and ETH-UCY Dataset, [14, 27] to demonstrate the effectiveness of our model. SDD consists of eight scenarios of the Stanford University campus, where the trajectories of various agents, such as pedestrians, bicyclists, and cars, are recorded. For long-term trajectory prediction, we use the popular setting of [20] to predict the future 30 seconds trajectory given the past 5 seconds trajectory from the long-term trajectory data in Stanford Drone Dataset. The ETH-UCY dataset contains five complex scenarios of crowd trajectories, where 1,536 pedestrians exhibit diverse interactions, such as collision avoidance, leading, following, and walking in groups. For short-term trajectory prediction, we use the ETH-UCY dataset and the Stanford Drone Dataset to predict the next 4.8 seconds trajectory given the previous 3.2 seconds trajectory.

Implementation Details We train our model with the Adam optimizer, batch size 4, learning rate 0.0001, on a single NVIDIA Tesla T4 GPU. The class of the scene semantic is set to 6. The pedestrian-scene encoder is implemented as a convolutional neural network consists of five convolutional layers with 32, 32, 64, 64, 64 channels. The goal decoder consists of a 5-layer convolutional network with channel numbers 64, 64, 64, 32, 32. The motion encoder and motion decoder are implemented with Multi-Layer Perceptrons(MLPs). The encoding dimension of the motion features is 32.The trajectory decoder has five convolutional layers with 66, 66, 66, 34, and 34 channels, respectively, to account for the additional two dimensions of the motion pattern and the goal feature. The scene layout modification map assigns a probability of 0.9 to frequently passed areas, 0.2 to occasionally passed areas, and 0 to impassable areas. The number of multimodal goals hyper-parameter is set to 20.

Baseline To evaluate the performance of our model, we benchmark it against several state-of-the-art methods. 1) SGAN [9] , a generative adversarial network that learns a distribution over plausible trajectories from a social context. 2) STGAT [10] , a sequence-to-sequence model that employs a graph attention network to capture the spatial and temporal dependencies among agents. 3) SGCN [33] , a method that uses a sparse directed graph to model the motion tendency and spatial interactions of agents, and a convolutional neural network to encode the graph features. 4) Trajectron++ [31], a modular, graph-structured recurrent model that incorporates agent dynamics and heterogeneous data, such as maps and scene semantics. 5) AgentFormer [43], a model that simultaneously models the time and social dimensions of trajectory prediction using a transformer architecture. 6) PECNet [21] , a model that first predicts the destinations of agents and then generates the whole trajectories conditioned on the destinations. 7) Y-net [20], a scene-compliant network that incorporates epistemic and aleatoric sources of uncertainty for long-term trajectory prediction. 8) GroupNet [38], a network models the interaction between agents with hypergraph. 9) MemoNet [39], a model applied the memorize mechanism to predict the future trajectory. 10) MID [7], a model utilizes the generative diffusion process to learn a distribution of future trajectory.

Evaluation Metrics We use two metrics to evaluate the performance of the models: Average Displacement Error (ADE) and Final Displacement Error (FDE). ADE is the average Euclidean distance between the predicted and the ground truth positions of the agents along the trajectory. FDE is the Euclidean distance between the predicted and the ground truth positions of the agents at the final time step.

4.1 Quantitative Evaluation

Long-term trajectory prediction We evaluate our model on Stanford Drone Dataset for long-term trajectory prediction task and compare the multimodal results with the baseline models.Table 1 presents the experiment result of our model and other baselines on the Stanford Drone Dataset for long-term trajectory prediction. We select the best of 20 multimodal trajectories generated from our model (following the common best-of-20 evaluation strategies) and other baseline models and calculate the ADE and FDE accordingly. We first acquire 20 multimodal goals and get the corresponding trajectory for each goal. For long-term trajectory prediction, we choose SGAN [9], PECNet [21] and Y-net [20] as baselines. Our GSMNet outperforms the previous best long-term model Y-net by 8.2% on ADE and 12.8% on FDE, which has not been improved for years. These results demonstrate that GSMNet can accurately predict long-term trajectories.

Short-term trajectory prediction We evaluate our model on Stanford Drone Dataset and ETH-UCY dataset for short-term trajectory prediction task and compare the multimodal results with the baseline models as in Table 2 and Table 3, respectively. For Stanford Drone Dataset, we choose SGAN [9], STGAT [10], Trajectron++ [31], PECNet [21], GroupNet [38], MemoNet [39]

Table 1: Quantitative experiment results (ADE/FDE) of previous state-of-the-art methods and our model for long-term trajectory prediction on Stanford Drone Dataset. We calculate the metrics for $T_{obs} = 5s$ and $T_{pre} = 30s$.

Method	ADE	FDE
SGAN	155.32	307.88
PECNet	72.22	118.13
Y − net	47.94	66.71
GSMNet	**44.03**	**58.17**

and MID [7] for comparison. For ETH-UCY, we contrast with the results of SGAN [9], STGAT [10], SGCN [33], Trajectron++ [31], AgentFormer [43] GroupNet [38], MemoNet [39], End-to-End [8] and MID [7]. Experimental results indicate that our method performance is competitive and outperforms the previous state-of-the-art methods in short-term prediction task with the lowest ADE and FDE.

Table 2: Quantitative experiment results (ADE/FDE) of previous state-of-the-art methods and our model for short-term trajectory prediction on Stanford Drone Dataset. We calculate the metrics for $T_{obs} = 3.2s$ and $T_{pre} = 4.8s$.

Method	ADE	FDE
SGAN	27.23	41.44
STGAT	14.23	26.67
Trajectron + +	9.90	16.80
PECNet	9.96	15.88
MID	9.73	15.32
MemoNet	8.56	**12.70**
GroupNet	9.31	16.11
End − to − End	8.60	13.90
GSMNet	**8.30**	**12.70**

4.2 Ablation Study

We conduct experiments to verify the effectiveness of the components proposed in our model, including the scene layout modification module, attention module, and motion pattern modeling module. We perform ablation experiments on the long-term trajectory prediction on the Stanford Drone Dataset, and the experimental results are shown in Table 4.

To prove the effectiveness of the scene layout modification module, we constructed variant 1, which samples the goals on the goal distribution without scene layout available constraints. Compared with variant 1, ADE and FDE

Table 3: Quantitative experiment results (ADE/FDE) of previous state-of-the-art methods and our model for short-term trajectory prediction on ETH-UCY. We calculate the metrics for $T_{obs} = 3.2s$ and $T_{pre} = 4.8s$.

Method	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
SGAN	0.81/1.52	0.72/1.61	0.60/1.26	0.34/0.69	0.42/0.84	0.58/1.18
STGAT	0.65/1.12	0.35/0.66	0.52/1.10	0.34/0.69	0.29/0.60	0.43/0.83
SGCN	0.63/1.03	0.32/0.55	0.37/0.70	0.29/0.53	0.25/0.45	0.37/0.65
Trajectron + +	0.43/0.86	0.12/0.19	**0.22**/**0.43**	**0.17**/0.32	**0.12**/0.25	0.20/0.39
AgentFormer	0.45/0.75	0.14/0.22	0.25/0.45	0.18/0.30	0.14/0.24	0.23/0.39
MemoNet	0.40/0.61	0.11/0.17	0.24/0.43	0.18/0.32	0.14/0.24	0.21/0.35
GroupNet	0.46/0.73	0.15/0.25	0.26/0.49	0.21/0.39	0.17/0.33	0.25/0.44
MID	0.39/0.66	0.13/0.22	0.22/0.45	0.17/**0.30**	0.13/0.27	0.21/0.38
GSMNet	**0.32**/**0.39**	**0.11**/**0.13**	0.25/**0.43**	0.18/0.28	0.15/0.26	**0.20**/**0.30**

Table 4: Ablation study on components of our proposed method for long-term trajectory prediction on Stanford Drone Dataset.

Variant	Description	Attention	Scene modify	Motion pattern	ADE/FDE
GSMNet	our whole model	+	+	+	**44.03/58.17**
1	w/o scene layout modify	+	-	+	45.67/64.74
2	w/o attention	-	+	+	45.96/60.96
3	w/o motion pattern	+	+	-	44.62/59.47

of GSMNet decreased by 3.6% and 10.1%, respectively, proving that the scene layout modification module can effectively avoid invalid prediction and improve the prediction accuracy of the endpoint and trajectory. To assess the role of the attention module, we constructed variant 2. This variant integrates the same pedestrian-scene correlation features in goal decoding and trajectory decoding. Compared with variant 2, ADE and FDE of GSMNet are decreased by 4.2% and 4.6%, respectively, proving that the attention module can effectively extract different information for different tasks. To verify the effect of the pedestrian movement pattern module, we designed variant 3. variant 3 is a model without a pedestrian motion pattern module, which combines the trajectory-based features and goals to generate a future trajectory. Compared with variant 3, ADE and FDE of the final model decreased by 1.3% and 2.2%, respectively. The results prove that motion pattern modeling is helpful to improve long-term trajectory prediction.

4.3 Qualitative Evaluation

Comparison of multimodal trajectories We visualize the predicted long-term trajectories on the scene image to verify that the predicted trajectories are consistent with the scene layout. We also compare our model with a state-of-the-art baseline model Y-net. Figure 3 shows three long-term pedestrian trajectories

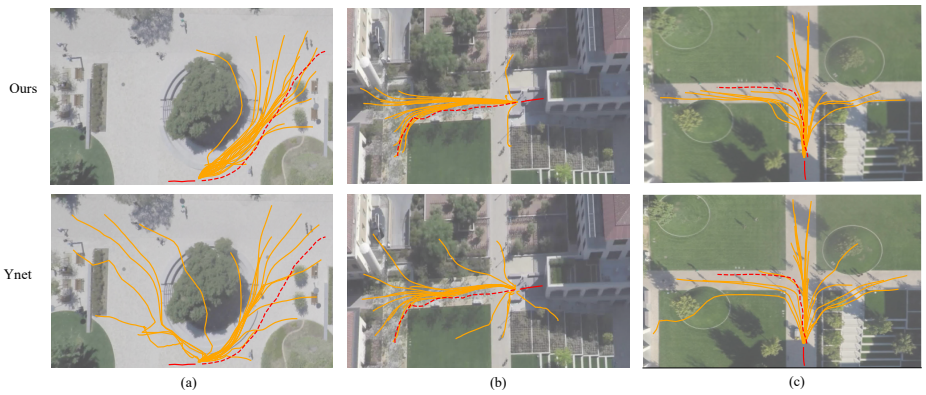

Fig. 3: Qualitative analysis of GSMNet and Y-net on the long-term trajectory prediction. Upper images are from GSMNet, lower images are from Y-net. (a) shows a square with a tree in the corner, (b) and (c) are intersections.

predicted by GSMNet and Y-net on different scenes. Figure 3(a) shows that the trajectories predicted by GSMNet avoid the tree in the center, and they follow both the scene layout and the motion pattern of pedestrians. However, the trajectory predicted by Y-net are more scattered, and some trajectories cross the central tree, which is unrealistic. Figure 3(b) shows that the trajectories predicted by GSMNet are distributed in the feasible regions of the scene. Most of the predicted trajectories tend to go straight as the real future trajectory and some of the predictions show the probability of turning at the intersection. Trajectories predicted by Y-net are more dispersed, and some trajectories go over the trees and the buildings. Figure 3(c) shows an intersection scene. GSMNet and Y-net can both predict that future trajectories have three different scenarios include going straight, turning left, and turning right. Some trajectories generated by Y-net are located on the lawn. Compared with Y-net, the trajectories predicted by GSMNet are more aligned with the pavement, which is reasonable.

Comparison of multimodal goals We evaluate the accuracy of goal prediction by comparing the predicted multimodal goal distribution of GSMNet and Y-net. Figure 4 shows that our model can predict more reasonable and realistic goals. In Figure 4 (a), the goals predicted by GSMNet are concentrated on the bottom of the road. However, the goals generated by Y-net are pointlessly scattered all over the scene. Figure 4 (b) shows that most of the goals predicted by GSMNet and Y-net are located on the road, but Y-net also predicts some goals on the trees, bushes, and other impassable regions, which are impossible for the agent to reach. In Figure 4 (c), our model predicts the goals that are aligned with the road, which is in accordance with the motion preference of the agent. But the goals predicted by Y-net are dispersed, and some of them are located in areas that are not easily accessible.

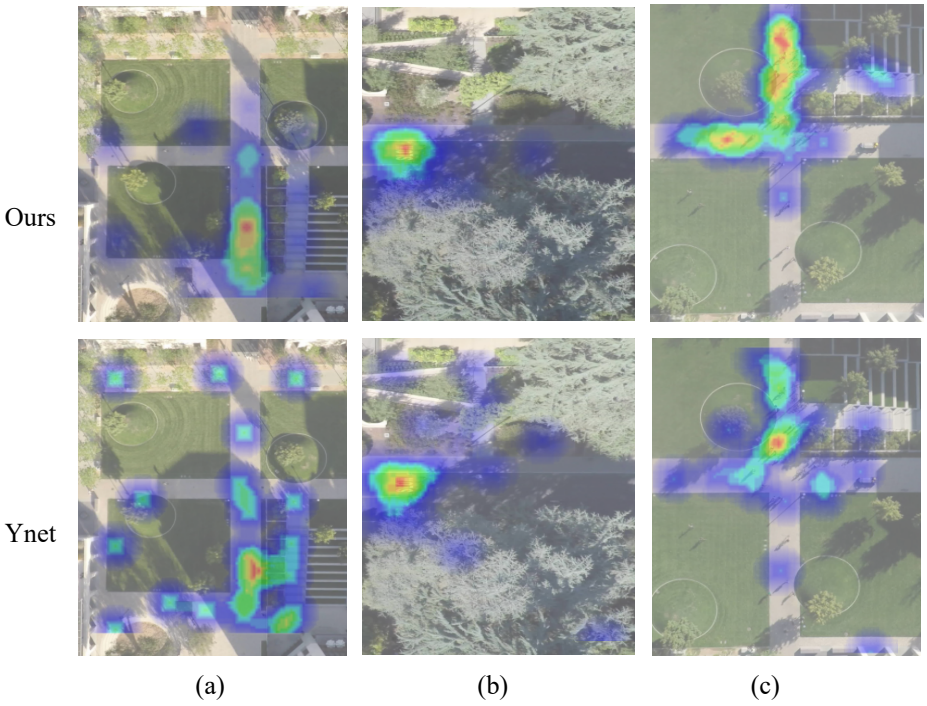

Fig. 4: Qualitative analysis of GSMNet and Y-net on goal distribution. Upper images are from GSMNet, lower images are from Y-net.

5 CONCLUSIONS

In this work, we present a novel model GSMNet for long-term trajectory prediction task that integrates multimodal goals, scene interaction and motion patterns. In this model, we first use historical trajectory and scene information to obtain the possible goals of the pedestrians. After that, we get the motion pattern from the historical trajectory. Finally, we combine the scene semantic information, motion pattern, and historical trajectory to generate the complete future trajectories. We propose the scene layout modification module to ensure the goals are in feasible areas of the scene and we use attention based U-Net to extract information correlation feature. We evaluate our model on two real-life datasets, Stanford Drone Dataset and ETH-UCY, for both long-term and short-term trajectory prediction. Our experimental results show that GSMNet outperforms existing state-of-the-art methods on the Stanford Drone Dataset for long-term trajectory prediction, and achieves competitive performance on both datasets for short-term trajectory prediction. We also visualize the predicted trajectories and goals on the scene, and proves that our model can generate more realistic and reasonable multimodal trajectories than previous methods.

Acknowledgments. This work was in part supported by the Major Program of National Natural Science Foundation of China (91938301).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social lstm: Human trajectory prediction in crowded spaces. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 961–971 (2016)
2. Bennewitz, Maren, Burgard, Wolfram, Cielniak, Grzegorz, Thrun, Sebastian: Learning motion patterns of people for compliant robot motion. International Journal of Robotics Research (2005)
3. Chandra, R., Guan, T., Panuganti, S., Mittal, T., Bhattacharya, U., Bera, A., Manocha, D.: Forecasting trajectory and behavior of road-agents using spectral clustering in graph-lstms. IEEE Robotics and Automation Letters **5**(3), 4882–4890 (2020)
4. Cheng, H., Liao, W., Tang, X., Yang, M.Y., Sester, M., Rosenhahn, B.: Exploring dynamic context for multi-path trajectory prediction. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 12795–12801. IEEE (2021)
5. Chiara, L.F., Coscia, P., Das, S., Calderara, S., Cucchiara, R., Ballan, L.: Goal-driven self-attentive recurrent networks for trajectory prediction (2022)
6. Gao, J., Sun, C., Zhao, H., Shen, Y., Anguelov, D., Li, C., Schmid, C.: Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11525–11533 (2020)
7. Gu, T., Chen, G., Li, J., Lin, C., Rao, Y., Zhou, J., Lu, J.: Stochastic trajectory prediction via motion indeterminacy diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17113–17122 (2022)
8. Guo, K., Liu, W., Pan, J.: End-to-end trajectory distribution prediction based on occupancy grid maps. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2232–2241 (2022). https://doi.org/10.1109/CVPR52688.2022.00228
9. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social gan: Socially acceptable trajectories with generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2255–2264 (2018)
10. Huang, Y., Bi, H., Li, Z., Mao, T., Wang, Z.: Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6272–6281 (2019)
11. Khan, M.S., Khan, M.I., Malik, S.U.R., Khalid, O., Azim, M., Javaid, N.: Matf: a multi-attribute trust framework for manets. EURASIP Journal on Wireless Communications and Networking **2016**(1), 1–17 (2016)
12. Kosaraju, V., Sadeghian, A., Martín-Martín, R., Reid, I., Rezatofighi, H., Savarese, S.: Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In: Advances in Neural Information Processing Systems. pp. 137–146 (2019)

13. Lee, N., Choi, W., Vernaza, P., Choy, C.B., Torr, P.H., Chandraker, M.: Desire: Distant future prediction in dynamic scenes with interacting agents. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 336–345 (2017)
14. Lerner, A., Chrysanthou, Y., Lischinski, D.: Crowds by example. In: Computer graphics forum. vol. 26, pp. 655–664. Wiley Online Library (2007)
15. Liang, J., Jiang, L., Murphy, K., Yu, T., Hauptmann, A.: The garden of forking paths: Towards multi-future trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10508–10518 (2020)
16. Lisotto, M., Coscia, P., Ballan, L.: Social and scene-aware trajectory prediction in crowded spaces. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019)
17. Liu, S., Wang, Y., Sun, J., Mao, T.: An efficient spatial-temporal model based on gated linear units for trajectory prediction. Neurocomputing (2021)
18. Ma, Y., Zhu, X., Zhang, S., Yang, R., Wang, W., Manocha, D.: Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 6120–6127 (2019)
19. Maeda, T., Ukita, N.: Fast inference and update of probabilistic density estimation on trajectory prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9795–9805 (2023)
20. Mangalam, K., An, Y., Girase, H., Malik, J.: From goals, waypoints & paths to long term human trajectory forecasting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15233–15242 (2021)
21. Mangalam, K., Girase, H., Agarwal, S., Lee, K.H., Adeli, E., Malik, J., Gaidon, A.: It is not the journey but the destination: Endpoint conditioned trajectory prediction. In: European Conference on Computer Vision. pp. 759–776. Springer (2020)
22. Mao, W., Xu, C., Zhu, Q., Chen, S., Wang, Y.: Leapfrog diffusion model for stochastic trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5517–5526 (2023)
23. Mohamed, A., Qian, K., Elhoseiny, M., Claudel, C.: Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14424–14432 (2020)
24. Monti, A., Bertugli, A., Calderara, S., Cucchiara, R.: Dag-net: Double attentive graph neural network for trajectory forecasting. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 2551–2558 (2021). https://doi.org/10.1109/ICPR48806.2021.9412114
25. Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.C., Lee, J.T., Mukherjee, S., Aggarwal, J., Lee, H., Davis, L., et al.: A large-scale benchmark dataset for event recognition in surveillance video. In: CVPR 2011. pp. 3153–3160. IEEE (2011)
26. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
27. Pellegrini, S., Ess, A., Schindler, K., Van Gool, L.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: 2009 IEEE 12th international conference on computer vision. pp. 261–268. IEEE (2009)
28. Ridel, D., Deo, N., Wolf, D., Trivedi, M.: Scene compliant trajectory forecast with agent-centric spatio-temporal grids. IEEE Robotics and Automation Letters **5**(2), 2816–2823 (2020)

29. Robicquet, A., Sadeghian, A., Alahi, A., Savarese, S.: Learning social etiquette: Human trajectory understanding in crowded scenes. In: European conference on computer vision. pp. 549–565. Springer (2016)
30. Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Rezatofighi, H., Savarese, S.: Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
31. Salzmann, T., Ivanovic, B., Chakravarty, P., Pavone, M.: Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In: Computer Vision – ECCV 2020. pp. 683–700 (2020)
32. Schöller, C., Knoll, A.: Flomo: Tractable motion prediction with normalizing flows. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7977–7984. IEEE (2021)
33. Shi, L., Wang, L., Long, C., Zhou, S., Zhou, M., Niu, Z., Hua, G.: Sgcn: Sparse graph convolution network for pedestrian trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8994–9003 (2021)
34. Shi, L., Wang, L., Zhou, S., Hua, G.: Trajectory unified transformer for pedestrian trajectory prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9675–9684 (October 2023)
35. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6479–6488 (2018)
36. Truong, X.T., Ngo, T.D.: Toward socially aware robot navigation in dynamic and crowded environments: A proactive social motion model. IEEE Transactions on Automation Science and Engineering **14**(4), 1743–1760 (2017)
37. Wang, C., Cai, S., Tan, G.: Graphtcn: Spatio-temporal interaction modeling for human trajectory prediction. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3450–3459 (2021)
38. Xu, C., Li, M., Ni, Z., Zhang, Y., Chen, S.: Groupnet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6498–6507 (2022)
39. Xu, C., Mao, W., Zhang, W., Chen, S.: Remember intentions: Retrospective-memory-based trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6488–6497 (2022)
40. Xue, H., Huynh, D.Q., Reynolds, M.: Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1186–1194. IEEE (2018)
41. Yao, Y., Atkins, E., Johnson-Roberson, M., Vasudevan, R., Du, X.: Bitrap: Bi-directional pedestrian trajectory prediction with multi-modal goal estimation. IEEE Robotics and Automation Letters **6**(2), 1463–1470 (2021)
42. Yu, C., Ma, X., Ren, J., Zhao, H., Yi, S.: Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In: European Conference on Computer Vision. pp. 507–523. Springer (2020)
43. Yuan, Y., Weng, X., Ou, Y., Kitani, K.M.: Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9813–9823 (2021)
44. Zhang, P., Ouyang, W., Zhang, P., Xue, J., Zheng, N.: Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12085–12094 (2019)