SocialMP: Learning Social Aware Motion Patterns via Additive Fusion for Pedestrian Trajectory Prediction

Tianci Gao[1] , **Yuzhen Zhang**[2] , **Hang Guo**[1] , **Pei Lv**[2]
[1]Henan Institute of Advanced Technology, Zhengzhou University
[2]School of Computer Science and Artificial Intelligence, Zhengzhou University
{gaotianci, zyzzhang, guohang}@gs.zzu.edu.cn, ielvpei@zzu.edu.cn

Abstract

Accurately capturing social interaction in complex scenarios is essential for pedestrian trajectory prediction task. The uncertainty in pedestrian interactions and the physical constraints imposed by the environment make this task challenging. To solve this problem, existing methods adopt dimensionality reduction algorithms to capture explainable human motions and behaviors. However, these approaches not only suffer from weak social awareness due to the inadequate feature extraction, but also overlook physical constraints, leading to predicted trajectories often cross unwalkable areas. To overcome these problems, we build an attention-based motion pattern representation, named SocialMP, which can effectively enhance the social awareness and environmental perception of motion patterns. Specifically, our method first characterizes the motion patterns through singular value decomposition and defines a visual field-based rule to model environmental social interaction. Then, an attention-based additive fusion mechanism is designed to enhance social awareness and environment perception of motion patterns. Therein, we integrate social interactions into motion patterns through cross-attention mechanism to generate latent motion patterns, and feed them into our devised additive fusion structure with backward connection for multiple iterations. Lastly, we design a map loss function by applying an additional penalty into average displacement error to prevent the pedestrians from passing through the unwalkable area. Extensive experiments on ETH-UCY and SDD datasets demonstrate that our SocialMP can not only improve prediction accuracy but also generate plausible trajectories.

1 Introduction

Perceiving, analyzing, and predicting future motion patterns of pedestrians are crucial for applications such as autonomous driving, intelligent transportation systems, robot navigation, and surveillance systems [Wang *et al.*, 2022; Huang *et al.*, 2023; Chen *et al.*, 2018; Quan *et al.*, 2021]. Given the 2D temporal coordinates of pedestrians within a scene, the task of trajectory prediction involves forecasting multiple plausible future trajectories that apply with social norms [Alahi *et al.*, 2016]. Traditional methods [Gupta *et al.*, 2018; Sadeghian *et al.*, 2019] directly infer the pedestrian's 2D coordinates based on past motion states which not only introduces uncertainty in the social interactions between pedestrians but also lacks interpretability.

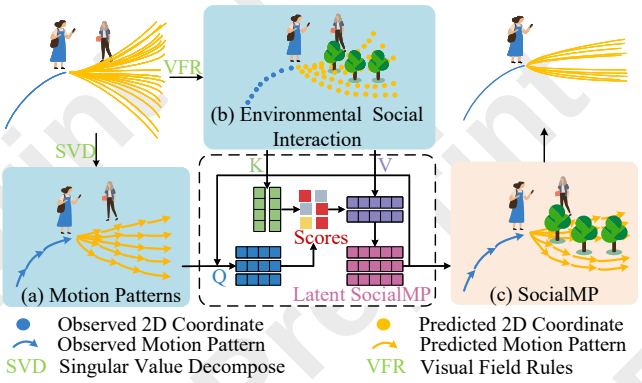

Figure 1: The process of pedestrians decreasing the uncertainty of their future trajectories from divergent (left top) to convergent (right top). (a) Avoiding neighbor pedestrian by intepretable motion patterns. (b) Perceiving the states of neighbor pedestrians and trees through unexplainable environmental social interaction. (c) Our approach integrate environmental social interaction into motion patterns via additive fusing mechanism to obtain SocialMP, which capture complex social interactions and explainable motion patterns.

To enhance the explainability of models, recent studies adopt dimensionality reduction algorithms to extract low-dimensional motion patterns [Hug *et al.*, 2020; Jazayeri and Jahangiri, 2022]. The key advantage of dimensionality reduction is its ability to capture interpretable motion patterns from available trajectories. For instance, considering some motion patterns of human, such as gradually slowing down to turn right or making a sharp turn while moving straight, some researchers introduce parametric curve functions [Huang *et al.*, 2019; Hug *et al.*, 2022] to capture explainable motion patterns. These approaches effectively reduce the dimensionality of trajectories by transforming sequential coordinates in the spatial domain into a condensed set of key points. Specif-

ically for the trajectory prediction task, Bae *et al.* [Bae *et al.*, 2023] employs a low-rank descriptor to approximate motion patterns from real-world trajectories. Nevertheless, this approach adopts singular value decompose algorithm which fails to model the complexity of interactions between pedestrians. Additionally, it overlooks the constraints imposed by the environment on pedestrian movement.

In this paper, departing from the current approaches that directly take dimensionality reduction approaches to represent interpretable motion patterns, we propose a novel framework to enhance the social awareness and environmental perception of motion patterns. The corresponding process of how pedestrians decrease the uncertainty of future trajectories from divergent to convergent is illustrated in Figure 1. The motion patterns of target pedestrian (blue clothes) are represented using dimensionality reduction approaches, which converges the predicted motion patterns in Figure 1 (a) to avoid the neighbor pedestrian. In contrary, Figure 1 (b) illustrates that the target pedestrian adopts visual field rules to model environmental social interaction which converges the predicted trajectory by perceiving the states of neighbor pedestrians and neighbor trees. Our model employs an attention-based additive fusion module integrating environmental social interaction into motion patterns to obtain social-aware motion patterns in Figure 1 (c) and we call that **SocialMP**. SocialMP can not only improve the prediction accuracy but also prevent the generation of future trajectory located in an unwalkable area. The main contributions are summarized as follows:

- We propose SocialMP representation for pedestrian trajectory prediction task which can effectively enhance the social awareness and environmental perception of motion patterns.
- We devise an attention-based additive fusion mechanism to capture the social awareness of motion patterns by integrating environmental social interaction into motion patterns.
- A map loss function is introduced to generate plausible trajectories by applying additional penalty to the average displacement error.
- Comprehensive experiments on two trajectory prediction datasets demonstrate that our approach achieves competitive results compared to the current state-of-the-art methods.

2 Related Work

Pedestrian trajectory prediction methods can be broadly categorized into model-based and model-free approaches [Jiangbei *et al.*, 2022].

Model-based Methods. These approaches use mathematical rules to model pedestrian behavior, often based on assumptions about human motion. The Social Force model [Helbing and Molnar, 1998] applies Newtonian laws to describe pedestrian movement. Other variations [Luber *et al.*, 2010; Pellegrini *et al.*, 2009b] enhance this model to better capture pedestrian interactions. Some models incorporate additional techniques, such as context-aware transfers [Xia *et al.*, 2022] or Neural Differential Equations [Jiangbei *et al.*, 2022], to explain pedestrian behavior. However, these methods struggle to capture the full complexity of real-world scenarios and require additional rules as system scale increases.

Model-free Methods. These methods focus on learning from data to model temporal and spatial relationships between pedestrians. Techniques such as social pooling [Alahi *et al.*, 2016] and generative models like CVAE [Ivanovic *et al.*, 2020; Xu *et al.*, 2024], GAN [Gupta *et al.*, 2018; Sadeghian *et al.*, 2019; Shilun *et al.*, 2021], Attention-based models [Saadatnejad *et al.*, 2024; Cheng *et al.*, 2023; Messaoud *et al.*, 2021; Zhang *et al.*, 2023] and diffusion-based models [Bae *et al.*, 2024; Mao *et al.*, 2023; Gu *et al.*, 2022] have been widely applied to capture spatio-temporal dependencies and pedestrian-environment interactions. While these models excel in data fitting, they often require specialized networks for multi-task prediction. Model-free approaches primarily emphasize fitting data through training deep neural networks, which usually lack the mathematical guidance that model-based methods inherently provide. To solve the above problem, Wong *et al.* [Wong *et al.*, 2024a] adopt an angle-based rule to model neighbor pedestrians' velocity, distance and direction. In this year, they also take the occupancy map into consideration to avoid the obstacles [Wong *et al.*, 2024b] . However, such methods consider all surrounding pedestrians and environmental information of the target pedestrian, which may lead to overfitting due to the limited field of view of humans. Unlike angle-based rules, to model environmental social interaction, we define visual field rules to simulate the real perspective of humans.

3 Methodology

Our method aims to enhance the social awareness of pedestrian motion patterns and reduce the uncertainty in pedestrian intentions. The corresponding pipeline is shown in Figure 2.

3.1 Problem Definition

The task of trajectory prediction is to forecast the future trajectories in a scene based on pedestrians' past trajectories. In this work, we denote the 2D trajectory coordinate as (x, y) and the scene map as M. Formally, we define N as the number of pedestrians in the scene and denote each pedestrian's history trajectory during T_h observed timesteps as $H_i = ((x_i^1, y_i^1), \cdots, (x_i^{T_h}, y_i^{T_h}))$. Correspondingly, the future ground truth trajectories are expressed as $F_i = ((x_i^{T_h+1}, y_i^{T_h+1}), \cdots, (x_i^{T_h+T_f}, y_i^{T_h+T_f}))$ during the future timesteps T_f. The trajectory prediction task can be formulated as follows: given the target pedestrian's past trajectories H_i and neighbor's past trajectories $\{H_1, \cdots, H_j, \cdots, H_N\}$ where $j \neq i$ in the scene map M, the objective is to predict s plausible future trajectories of the target pedestrian, denoted as $\hat{F}_i$.

3.2 Environmental Social Interaction

The extraction process of environmental social interaction is mainly divided into three steps. Firstly, we convert the scene segmentation map into a 100×100 pixel-level map. By determining the infeasible regions based on the pixel values of the

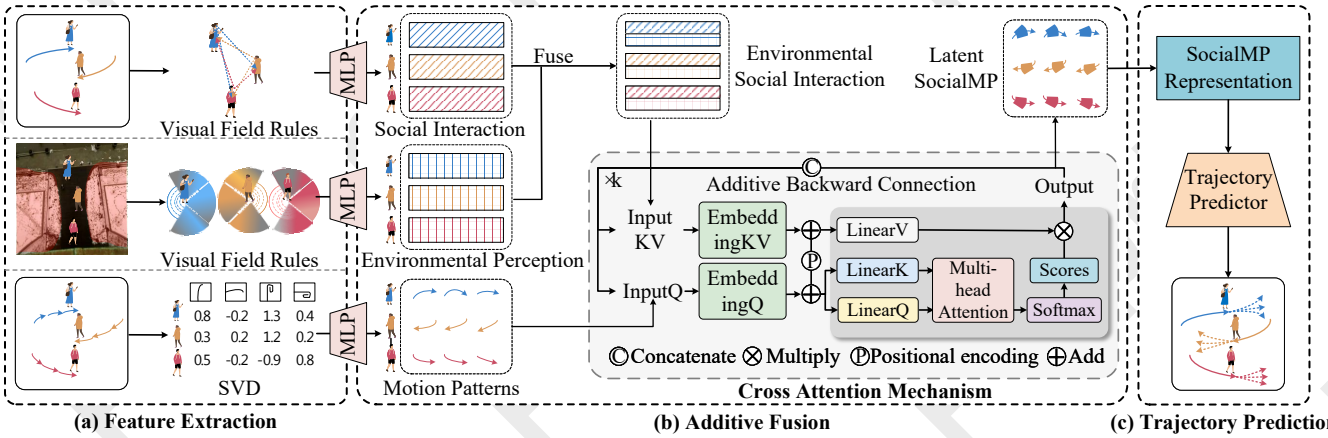

Figure 2: The overall framework of SocialMP representation. (a) Feature Extraction module extracts different dimensional features using corresponding methods. The Visual Field Rules are used to extract social interactions among humans (H-H) and environmental perception between human and environment (H-E). The Singular Value Decompose approach is used to capture explainable motion patterns. (b) The Additive Fusion module integrates environmental social interactions into motion patterns through an iterative cross-attention mechanism. (c) The Trajectory Prediction module incorporates the SocialMP representation into the backbone prediction models to generate accurate and socially-aware multi-modal predicted trajectories.

segmentation map (0 or 255), we extract the coordinates of the infeasible areas. These coordinates are then transformed into real-world coordinates using a homography matrix, and these coordinates are treated as stationary pedestrians.

Next, we define visual field rules to model environmental social interaction. Treue [Treue, 2003] proposes the concept of visual attention and points out that humans exhibit varying degrees of attention from different perspectives. Inspired by these, we consider the limited field of view and attention span of humans and divide the surrounding area of the target pedestrian into three zones: (1) None Attention. This indicates that the pedestrian or obstacle is behind the target pedestrian and cannot be noticed. (2) Weak Attention. This indicates that the pedestrian or obstacle is in the side-front of the target pedestrian, which can be noticed through peripheral vision, but it has minimal impact on the pedestrian. (3) Strong Attention. This refers to direct attention that the pedestrian can consciously perceive, and it exerts a strong interactive influence. Based on these, We set the interaction impact coefficients α, β, and γ to constrain the influence. Specifically, we calculate the relative angle $rel(i, j)$ between the direction of travel of the target pedestrian i and the surrounding pedestrian j. If the value is within the range from $\frac{2}{3}\pi$ to $\frac{4}{3}\pi$, the target pedestrian is defined as None Attention, with a value of α; if the value is within the range from $\frac{1}{3}\pi$ to $\frac{2}{3}\pi$ or range from $\frac{4}{3}\pi$ to $\frac{5}{3}\pi$, the pedestrian is defined as Weak Attention, with a value of β; if the value is between 0 and $\frac{1}{3}\pi$ or between $\frac{5}{3}\pi$ and 2π, the pedestrian is defined as Strong Attention, with a value of γ. The values of α, β, and γ are hyperparameters that can be adjusted to adapt to different scenarios. In this case, we set $\alpha = 0$, $\beta = 0.5$, and $\gamma = 1$.

Lastly, we use visual field rules to model environmental social interaction. Given the trajectories of N pedestrians (including stationary pedestrians), the interactions between the target pedestrian and its neighbors are extracted by predefined rules. We use an angle-based approach to capture the relative motion of each pedestrian. The angle θ_{ij} between pedestrian j and target i is computed as:

$$\theta_{ij} = \arctan\left(\frac{y_j^{T_h} - y_i^{T_h}}{x_j^{T_h} - x_i^{T_h}}\right), \tag{1}$$

where θ_{ij} represents the relative direction of pedestrian j to pedestrian i. The representations of velocity f_{vlc}^i, direction f_{drt}^i, and distance f_{dst}^i are defined as:

$$\begin{aligned} f_{vlc}^i &= \frac{1}{N}\sum_{j=1}^{N}\left\|(x_j^{T_h}, y_j^{T_h}) - (x_j^1, y_j^1)\right\|_2, \\ f_{drt}^i &= \frac{1}{N}\sum_{j=1}^{N}\theta_{ij}, \\ f_{dst}^i &= \frac{1}{N}\sum_{j=1}^{N}\left\|(x_j^{T_h}, y_j^{T_h}) - (x_i^{T_h}, y_i^{T_h})\right\|_2. \end{aligned} \tag{2}$$

Since the target pedestrian can not simultaneously focus on every surrounding pedestrian in all directions, we divide the vicinity of the target pedestrian into p angular partitions, and calculate the social interaction representations $f_{nei}^{p,i}$ relative to the target pedestrian i in the direction of p. It is important to note that if no pedestrians are present within a specific directional area, the corresponding factors will be set to zero. Each partition's representation $f_{nei}^{p,i}$ is computed using the following formula:

$$f_{nei}^{p,i} = \text{Concat}(f_{vlc}^{p,i}, f_{drt}^{p,i}, f_{dst}^{p,i}). \tag{3}$$

The social interaction features f_{si}^i for the target pedestrian i are computed by concatenating the representations from multiple angular partitions:

$$f_{si}^i = \text{Concat}(f_{nei}^{1,i}, f_{nei}^{2,i}, \ldots, f_{nei}^{p,i}). \tag{4}$$

3.3 Motion Pattern Descriptor

We extract interpretable motion patterns by following Eigentrajectory [Bae *et al.*, 2023], which uses singular value decomposition to represent motion patterns. The extraction process aligns with that described in the original work.

3.4 Additive Fusion

In the additive fusion module, we first embed the temporal observed trajectories into the original environmental social interaction f_{si} and motion patterns f_m, enabling the fuser to extract spatio-temporal dependencies. Then an additive cross-attention mechanism is adopted to compute attention scores between environmental social interaction and motion patterns and result in social-awared motion patterns.

Temporal Embedding. Environmental social interaction capture the spatial interactive features at time step t_h, but do not inherently account for the sequence of observed trajectories. To address this, we embed the temporal observed trajectories into the social representation f_{sr} to capture temporal dependencies. Motion patterns consist of r singular vectors obtained from the spatio-temporal data decomposition of a scene, encapsulating spatio-temporal information. We also embed the initial positions of pedestrians into the motion patterns to retain information about their initial positions. Formally:

$$\begin{aligned} X_s^i &= \text{MLP}(\tanh(\text{Concat}(\text{MLP}(f_{si}^i), \text{MLP}(H^i)))), \\ X_m^i &= \text{MLP}(\tanh(\text{Concat}(f_m^i, H^i[0]))), \end{aligned} \tag{5}$$

where X_s^i represents the temporal environmental social interaction with a shape of (N, T, d_s) and X_m^i denotes the temporal motion patterns with a shape of (N, T, d_m). H^i and $H^i[0]$ represent the observed trajectories and the original position of pedestrian i, respectively.

Cross Attention Mechanism. An attention-based mechanism is employed to integrate high-dimensional environmental social interaction f_{si}^i into low-dimensional motion patterns f_m^i. We utilize cross attention to compute attention scores between environmental social interaction and motion patterns. The formula for calculating the attention scores of X_m on X_s is as follows:

$$\text{CA}(X_m, X_s) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_s}}\right)V, \tag{6}$$

where $Q = X_m W_Q$ and $K = V = X_s W_K$. Through the multi-head cross-attention mechanism, a fused representation SF^i is obtained and concatenated with the original representation. Mathematically,

$$\begin{aligned} X_m^{i,k} &= \text{MLP}\left(\text{Concat}(\text{CA}(X_m^{i,k-1}, X_s^{i,k-1}), X_m^{i,k-1})\right), \\ X_s^{i,k} &= \text{MLP}\left(\text{Concat}(\text{CA}(X_m^{i,k-1}, X_s^{i,k-1}), X_s^{i,k-1})\right), \end{aligned} \tag{7}$$

where k is a hyperparameter and we empirically set k to 6.

3.5 Loss Function

SocialMP uses additional loss functions L_{smp} and L_{map} to train the baseline models. Specifically, to enhance the plausibility of the predicted endpoint, we first convert the predicted real-world coordinates of pedestrians into pixel-level coordinates on the segmentation map using the homography matrix provided by the dataset. Then we determine whether the predicted coordinate points lie within unwalkable areas by examining the pixel values of the scene segmentation map. This process is expressed as:

$$Map(x, y) = \begin{cases} 1, & \text{if } (x,y) \text{ in the obstacle area,} \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

while the final predicted endpoint falls within unwalkable area, a penalty is applied to the model's loss function. The map loss function is defined as follows:

$$\begin{aligned} L_{map} = \text{Map}(x_n^{T_w}, y_n^{T_w}) \\ \cdot \frac{1}{N}\sum_{n=1}^{N} \left\| (\hat{x}_n^{T_w}, \hat{y}_n^{T_w}) - (x_n^{T_w}, y_n^{T_w}) \right\|_2, \end{aligned} \tag{9}$$

here, we use final displacement error for L_{map} to provide an additional penalty when the model predicts an unreasonable endpoint.

For the loss function of motion patterns, we use its coefficient c to measure the degree of deviation, which is calculated as shown in the following equation:

$$L_{smp} = \frac{1}{N}\sum_{n=1}^{N} \|\hat{c}_{f,n} - c_{f,n}\|. \tag{10}$$

The final loss is a combination of two loss functions. Formally,

$$Loss = \lambda_1 L_{smp} + \lambda_2 L_{map}, \tag{11}$$

where we empirically set λ_1 and λ_2 to 1.

4 Experiments

4.1 Experimental Setup

Datasets. The ETH [Pellegrini *et al.*, 2009a] and UCY [Lerner *et al.*, 2007] (ETH-UCY) dataset covers 1523 pedestrians' trajectories across five scenes: ETH, Hotel, Univ, Zara1 and Zara2. These trajectories are based on real-world coordinates captured by an aerial surveillance camera. Similar to the method [Gupta *et al.*, 2018], we adopt the standard leave-one-out strategy for the training and evaluation. Stanford Drone Dataset (SDD) [Robicquet *et al.*, 2016] has 60 drone videos, including 5243 pedestrians in eight different university campus scenes. Following previous work [Liang *et al.*, 2020], we split 60% videos to train, 20% to validate, and 20% to test in pixel coordinates.

Metrics. We use *minADE*$_{20}$*/minFDE*$_{20}$ [Alahi *et al.*, 2016; Gupta *et al.*, 2018] metrics that the best Average/Final Displacement Error (*ADE/FDE*) over 20 generated trajectories to measure the prediction accuracy and endpoint accuracy respectively. Simultaneously, we also introduce the metric *TCC/COL* [Tao *et al.*, 2020; Liu *et al.*, 2021], therein The Temporal Correlation Coefficient measures the Pearson correlation coefficient of motion patterns, and the Collision Rate calculates the percentage of collision cases between pedestrians on the predicted trajectories.

Dataset	SGCN/SMP-SGCN				STGCNN/SMP-STGCNN				Implicit/SMP-Implicit			
	ADE↓	FDE↓	TCC↑	COL↓	ADE↓	FDE↓	TCC↑	COL↓	ADE↓	FDE↓	TCC↑	COL↓
ETH	0.57/**0.35**	1.00/**0.55**	**0.55**/0.51	1.69/**1.16**	0.65/**0.34**	1.10/**0.57**	**0.51**/0.47	1.80/**1.38**	0.66/**0.34**	1.44/**0.55**	**0.53**/0.45	1.93/**1.22**
HOTEL	0.31/**0.12**	0.53/**0.19**	**0.30**/0.27	2.52/**1.49**	0.50/**0.14**	0.86/**0.23**	0.27/**0.28**	3.94/**1.58**	0.20/**0.13**	0.36/**0.20**	**0.29**/0.28	3.67/**1.43**
UNIV	0.37/**0.24**	0.67/**0.39**	**0.69**/0.64	**6.85**/7.33	0.44/**0.23**	0.80/**0.39**	0.64/**0.74**	9.69/**8.71**	0.31/**0.22**	0.60/**0.39**	0.66/**0.83**	**7.74**/8.58
ZARA1	0.29/**0.21**	0.51/**0.33**	0.75/**0.83**	**0.79**/1.24	0.34/**0.22**	0.53/**0.39**	0.71/**0.82**	2.53/**1.29**	0.25/**0.21**	0.50/**0.39**	0.71/**0.82**	2.38/**1.27**
ZARA2	0.23/**0.14**	0.42/**0.25**	0.49/**0.72**	**2.23**/6.09	0.31/**0.16**	0.48/**0.28**	0.39/**0.68**	7.15/**5.99**	0.22/**0.15**	0.43/**0.26**	0.47/**0.75**	**5.48**/5.58
AVG	0.35/**0.21**	0.63/**0.34**	0.55/**0.59**	**2.82**/3.46	0.45/**0.22**	0.75/**0.37**	0.50/**0.60**	5.02/**3.79**	0.33/**0.21**	0.67/**0.36**	0.53/**0.63**	4.17/**3.62**
SDD	11.42/**7.83**	18.89/**13.26**	0.57/**0.66**	4.45/**1.26**	20.76/**8.54**	33.18/**13.74**	0.47/**0.59**	0.68/**0.46**	15.74/**9.26**	23.15/**13.47**	0.55/**1.75**	**1.64**/2.33

Table 1: Comparisons between baseline models and the corresponding models with the designed SocialMP (SMP-based) through $T_h = 8$ frames of observations to predict future $T_f = 12$ frames of trajectories. **Bold**: Best.

Method	SocialGAN	SOPHIE	Pecnet	BCDiff	Graph-TERN	MRL	SMEMO	SMP-SGCN(Ours)
ADE↓	27.23	16.27	9.96	9.05	8.43	8.22	<u>8.11</u>	**7.83**
FDE↓	41.44	29.38	15.88	14.86	14.26	13.39	**13.06**	<u>13.26</u>

Table 2: Comparisons with current state-of-the-art methods on SDD. **Bold**: Best. <u>Underline</u>: Second Best.

Method	Metrics			
	ADE↓	FDE↓	TCC↑	COL↓
Social-LSTM	0.72	1.54	0.21	6.74
Social-GAN	0.61	1.21	0.37	5.43
PECNet	0.54	0.87	0.43	6.42
Trajectron++	0.31	0.52	0.36	5.42
STGAT	0.31	0.62	0.44	<u>2.43</u>
AgentFormer	0.23	0.39	0.41	4.34
GroupNet	0.25	0.44	0.58	**2.12**
GP-Graph	0.23	0.39	0.51	3.24
Graph-TERN	0.24	0.38	0.45	3.23
SMEMO	<u>0.22</u>	<u>0.35</u>	0.46	4.32
SGCN	0.35	0.63	0.55	2.82
STGCNN	0.45	0.75	0.50	5.02
Implicit	0.33	0.67	0.53	4.17
SMP-SGCN (Ours)	**0.21**	**0.34**	0.59	3.73
SMP-STGCNN (Ours)	<u>0.22</u>	0.37	<u>0.60</u>	3.79
SMP-Implicit (Ours)	**0.21**	0.36	**0.63**	3.62

Table 3: Comparisons with the current state-of-the-art methods on ETH-UCY dataset. **Bold**: Best. <u>Underline</u>: Second Best.

Baseline models. We evaluate our approach against common baselines such as SGCN [Shi *et al.*, 2021], STGCNN [Mohamed *et al.*, 2020], and Implicit [Mohamed *et al.*, 2022]. Other baseline methods are used to compare with the state-of-the-art (SOTA) performance.

4.2 Quantitative Analyses

Performance Evaluation of Representations. We evaluate some original models and the corresponding models (SMP-based) with the SocialMP on ETH-UCY and SDD. As illustrated in Table 1, SMP-based models decrease the values of *ADE*, *FDE*, *COL* and increase the value of *TCC* in most cases which demonstrates that the SocialMP representations exhibit superior performance compared to the baseline models. In particular, the results of *ADE* and *FDE* decrease by at least 36.2% and 45.2% compared to the model without representations. In terms of reliability measures, except for a slight performance degradation in *COL*, all other metrics' results are improved on the ETH-UCY and SDD datasets. The experimental results demonstrate that our SocialMP can effectively integrate social awareness into motion patterns to handle different scenarios.

Comparisons with State-of-the-Art Models. We compare some models inserting SocialMP with the state-of-the-art methods on SDD and ETH-UCY respectively. Table 2 illustrates that SGCN with SoicalMP (SMP-SGCN) obtains competitive performance on SDD. In detail, SMP-SGCN achieves impressive predictive performance with 3.5% better *ADE* compared with SMEMO and has a performance decrease of only 1.5% in the FDE compared to the SMEMO model. In Table 3, compared with SMEMO, SMP-SGCN achieves better performance with 4.5% better *ADE*, 2.9% better *FDE*, 22.0% better *TCC* and 13.7% better *COL*. Specifically, SMP-SGCN achieves state-of-the-art performance on the *ADE* and *FDE* metrics and SMP-Implicit reached state-of-the-art results on the *TCC* metric. The performance on the *COL* of SMP-based models has slightly declined compared to these later methods, and the reason is that SocialMP emphasizes collision avoidance between pedestrians and obstacles, overlooking potential collisions between pedestrians. These results demonstrate that SocialMP-based models can achieve competitive results, and further reflect that the SocialMP can extract more effective representation of motion patterns.

4.3 Qualitative Analyses

Unimodal trajectories of all pedestrians on the scene of UNIV. In Figure 3, we visualize the effect of predicting all pedestrians simultaneously. Visual Field SMP-SGCN utilizes the hyperparameters α, β, and γ to adjust the intensity of attention to social interactions, controlling the degree of influence through three types of attention: none, weak, and strong. Pedestrians with none attention (gray lines) do not affect the motion patterns of the target pedestrian. Weak attention (orange lines) slightly influences the motion patterns of the target pedestrian, enabling the model to better simulate pedestrian movement without disrupting the target pedestrian's trajectory. Strong attention (red lines) places greater emphasis on social interactions, effectively preventing conflicts between the target pedestrian and neighboring pedestrians, resulting in more coordinated trajectories. By flexibly adjusting

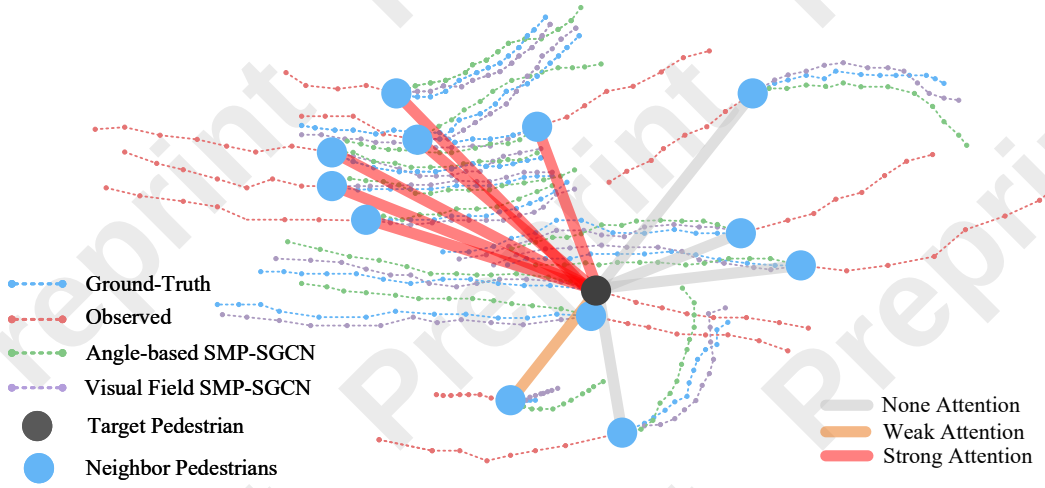

Figure 3: The visualization of predicted trajectories of visual field SMP-SGCN (Ours) and angle-based SMP-SGCN (original) in the crowed UNIV scene.

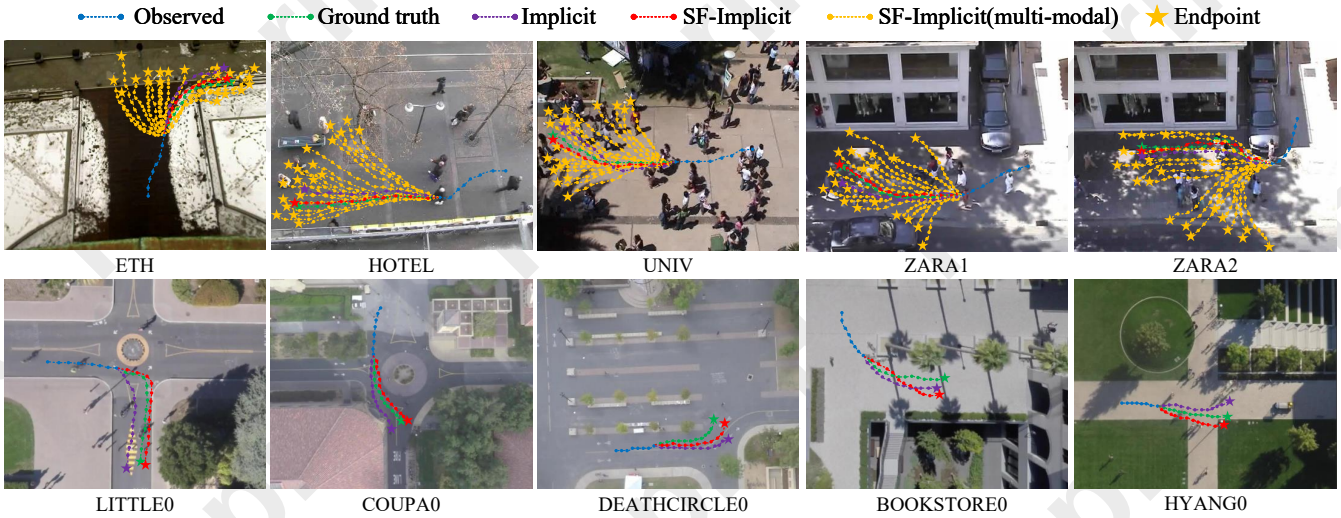

Figure 4: The visualization of predicted trajectories of SMP-Implicit (Ours) and Implicit (original) in ETH-UCY and SDD scenes. The first row of five scenes represents the multi-modal visualization of five sub-scenes from ETH-UCY, while the second row of five scenes shows the best of 20 trajectories visualization from SDD.

the weight of social interactions, Visual Field SMP-SGCN effectively avoids overfitting, captures the dynamic nature of social interactions, and produces predictions that are closer to the ground truth trajectory.

Multi-modal trajectories on ETH-UCY. In the top row of Figure 4, we visualize the predicted multi-modal trajectories from five sub-scenes on ETH-UCY dataset respectively, which involves different scenes with sparse crowds or dense crowds including turning right and going straight. The SMP-Implicit predicts reasonable and accurate trajectories. Particularly, in the UNIV scene with high-density crowds, SMP-Implicit learns the motion patterns of slowly turning right, while the motion patterns with quickly turning right predicted by Implicit deviate from the ground truth trajectory. In the ZARA1 and ZARA2 scenes, the multi-modal trajectories of target pedestrian do not pass through the building, cars and other unwalkable area. The reason is that the surrounding obstacle information is integrated into the SocialMP representations. These results demonstrated that our SocialMP representations can effectively handle complex scenarios and prevent the pedestrians from passing through the unwalkable area.

Unimodal trajectories on SDD. The unimodal trajectories on SDD presented in Figure 4 intuitively reflects the accuracy of SMP-Implicit. We can see that SMP-Implicit performs better accuracy than the Implicit. Specifically, in the scene of BOOKSTORE0, the predictions of SMP-Implicit initially deviate from the ground truth trajectory in order to avoid the trees. Then, the pedestrian adjusts the motion pattern to walk toward the final position and achieve better accuracy com-

Features	SGCN				STGCNN				Implicit			
	ADE↓	FDE↓	TCC↑	COL↓	ADE↓	FDE↓	TCC↑	COL↓	ADE↓	FDE↓	TCC↑	COL↓
MP	0.218	0.362	0.590	3.663	0.229	0.383	0.589	6.212	0.224	0.372	0.619	5.432
SI	0.242	0.476	0.453	4.474	0.263	0.394	**0.637**	6.531	0.325	0.379	0.554	6.430
SI+MP	0.227	0.353	0.539	**3.213**	0.232	0.386	0.573	4.683	0.227	0.363	0.563	5.437
SI+SO	0.236	0.349	0.572	3.769	0.236	0.381	0.627	5.898	0.232	0.371	**0.674**	3.865
MP+SO	0.238	0.357	0.563	3.652	0.246	0.379	0.512	4.336	0.219	0.368	0.474	5.476
SI+SO+MP(Ours)	**0.213**	**0.343**	**0.593**	3.733	**0.218**	**0.373**	0.597	**3.790**	**0.210**	**0.359**	0.625	**3.621**
Gain(%)	+2.3%	+1.4%	+0.5%	-16.2%	+4.8%	+1.8%	-6.3%	+12.6%	+4.1%	+1.1%	-7.3%	+3.7%

Table 4: Ablation studies on the components of SocialMP on the ETH-UCY dataset. "SI", "SO" and "MP" indicate whether social interactions, surrounding obstacles and motion patterns are included in SocialMP. **Bold**: Best.

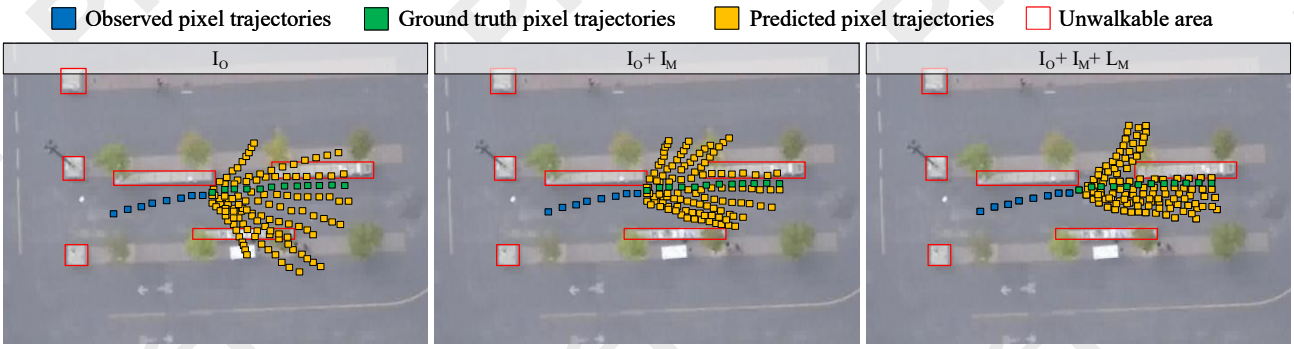

Figure 5: The visualization of the impact of map loss function on generated trajectories with SMP-SGCN in BOOKSTORE.

pared to the Implicit method.

4.4 Ablation Studies

Effectiveness of each component on SocialMP. We validate the effectiveness of each component of SocialMP and the quantification results on ETH-UCY are listed in Table 4. Therein, we compare "SI+SO+MP" (SocialMP) with its individual ("MP","SI") and combined components ("SI+MP", "SI+SO", "MP+SO") features on ETH-UCY datasets, where "Gain" represents the performance improvement results between the "SI+SO+MP" and the best of other features. Comparing "MP" with "SI+SO", we can see that using motion patterns as trajectory features generally yields better performance. This comparison also indicates that motion patterns have a superior representational expression compared to environmental social interaction. When comparing "SI+SO+MP" to "MP", SocialMP achieves an average improvement of 2.73% and 1.43% on *ADE* and *FDE* compared with MP-based models. Though, such improvements come at the price of a decline in part *TCC* and *COL* performances. The reason is that the model's training loss function is primarily depended on *ADE* and *FDE*, without incorporating *TCC* or *COL* metrics into the loss function. These results demonstrate that although individual environmental social interaction yields worse performance than solely relying on motion patterns for predictions, integrating environmental social interaction into motion patterns can still enhance the model's performance.

Validation of map loss function. To validate the effectiveness of map loss function, we visualized predicted trajectories of SMP-SGCN in the BOOKSTORE scene. As shown in Figure 5, in the first column, I_O solely applies pedestrians' coordinates without map and the predicted trajectories pass through the unwalkable area. In the middle column, $I_O + I_M$ takes map I_M into consideration based on I_O model and the predicted trajectory gradually moves away from the unwalkable area. There still has a small portion of the predicted trajectory lies within the unwalkable area, and the reason is that the loss function encourages the model to generate trajectories that are as close as possible to the ground truth trajectory. In the last column, $I_O + I_M + L_M$ incorporates an extra map loss function L_M to impose a penalty when the predictions of SMP-SGCN pass through the unwalkable area. The generated motion patterns of turning left are away from the unwalkable area, which demonstrates that the effectiveness of proposed map loss function.

5 Conclusion

In this work, we present SocialMP, an innovative representation for pedestrian trajectory prediction aiming to reduce the uncertainty in pedestrian interactions and enhance environmental perception ability. By incorporating environmental social interactions into motion patterns via an attention-based additive fusion mechanism, SocialMP effectively models complex social behaviors and accounts for physical constraints. Additionally, the proposed map loss function ensures more plausible predictions by penalizing those trajectories that pass through unwalkable areas. Future research will explore to leverage multi-modal data, such as LiDAR or point clouds, to further enhance prediction accuracy.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62372415, and Outstanding Youth Science Fund of Henan Province under Grant 242300421050.

Contribution Statement

Tianci Gao and Yuzhen Zhang contributed equally to this work and are designated as co-first authors. Hang Guo provided critical review of the manuscript. Pei Lv served as the corresponding author and is responsible for all communications related to this manuscript.

References

[Alahi *et al.*, 2016] 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, Las Vegas, 2016. IEEE Computer Society.

[Bae *et al.*, 2023] Inhwan Bae, Jean Oh, and Hae-Gon Jeon. Eigentrajectory: Low-rank descriptors for multi-modal trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9983–9995, Los Alamitos, 2023. IEEE Computer Society.

[Bae *et al.*, 2024] Inhwan Bae, Young-Jae Park, and Hae-Gon Jeon. Singulartrajectory: Universal trajectory predictor using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle WA, USA, 2024. IEEE Computer Society.

[Chen *et al.*, 2018] Zhixian Chen, Chao Song, Yuanyuan Yang, Baoliang Zhao, Ying Hu, Shoubin Liu, and Jianwei Zhang. Robot navigation based on human trajectory prediction and multiple travel modes. *Applied Sciences*, 8(2205), 2018.

[Cheng *et al.*, 2023] Hao Cheng, Mengmeng Liu, Lin Chen, Hellward Broszio, Monika Sester, and Michael Ying Yang. Gatraj: A graph- and attention-based multi-agent trajectory prediction model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 205:163–175, 2023.

[Gu *et al.*, 2022] Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, and Jiwen Lu. Stochastic trajectory prediction via motion indeterminacy diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17113–17122, New Orleans, LA, USA, 2022. IEEE Computer Society.

[Gupta *et al.*, 2018] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. 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, Los Alamitos, CA, USA, 2018. IEEE Computer Society.

[Helbing and Molnar, 1998] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical Review E*, 51, 1998.

[Huang *et al.*, 2019] 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, Seoul, Korea, 2019. IEEE Computer Society.

[Huang *et al.*, 2023] Z. Huang, H. Liu, and C. Lv. Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3880–3890, Los Alamitos, CA, USA, 2023. IEEE Computer Society.

[Hug *et al.*, 2020] Ronny Hug, Wolfgang Hübner, and Michael Arens. Introducing probabilistic bézier curves for n-step sequence prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(06):10162–10169, 2020.

[Hug *et al.*, 2022] Ronny Hug, Stefan Becker, Wolfgang Hübner, Michael Arens, and Jürgen Beyerer. Bézier curve gaussian processes, May 2022.

[Ivanovic *et al.*, 2020] Karen Ivanovic, B.and Leung, Edward Schmerling, and Marco Pavone. Multimodal deep generative models for trajectory prediction: A conditional variational autoencoder approach. *IEEE Robotics and Automation Letters*, 6:295–302, 2020.

[Jazayeri and Jahangiri, 2022] Mohammad Sadegh Jazayeri and Arash Jahangiri. Utilizing b-spline curves and neural networks for vehicle trajectory prediction in an inverse reinforcement learning framework. *Journal of Sensor and Actuator Networks*, 11(1), 2022.

[Jiangbei *et al.*, 2022] Yue Jiangbei, Manocha Dinesh, and Wang He. Human trajectory prediction via neural social physics. In *European Conference on Computer Vision*, pages 376–394, Berlin, Heidelberg, 2022. Springer-Verlag.

[Lerner *et al.*, 2007] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. *Comput. Graph. Forum*, 26:655–664, 2007.

[Liang *et al.*, 2020] Junwei Liang, Lu Jiang, and Alexander Hauptmann. Simaug: Learning robust representations from simulation for trajectory prediction. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII*, page 275–292, Glasgow, United Kingdom, 2020. Springer-Verlag.

[Liu *et al.*, 2021] Yuejiang Liu, Qi Yan, and Alexandre Alahi. Social nce: Contrastive learning of socially-aware motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15118–15129, Montreal, QC, Canada, 2021. Piscataway, NJ: IEEE.

[Luber *et al.*, 2010] Matthias Luber, Johannes Stork, Gian Diego Tipaldi, and Kai Arras. People tracking with human motion predictions from social forces. In *2010 IEEE international conference on robotics and automation*, pages 464–469, Anchorage, Alaska, USA, 2010. Curran Associates, Inc.

[Mao *et al.*, 2023] W. Mao, C. Xu, Q. Zhu, S. Chen, and Y. Wang. Leapfrog diffusion model for stochastic trajectory prediction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5517–5526, Los Alamitos, CA, USA, jun 2023. IEEE Computer Society.

[Messaoud *et al.*, 2021] Kaouther Messaoud, Itheri Yahiaoui, Anne Verroust-Blondet, and Fawzi Nashashibi. Attention based vehicle trajectory prediction. *IEEE Transactions on Intelligent Vehicles*, 6(1):175–185, 2021.

[Mohamed *et al.*, 2020] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14412–14420, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society.

[Mohamed *et al.*, 2022] Abduallah Mohamed, Deyao Zhu, Warren Vu, Mohamed Elhoseiny, and Christian Claudel. Social-implicit: Rethinking trajectory prediction evaluation and the effectiveness of implicit maximum likelihood estimation. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, page 463–479, Tel Aviv, Israel, 2022. Springer-Verlag.

[Pellegrini *et al.*, 2009a] S. Pellegrini, A. Ess, K. Schindler, and L. 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, Kyoto, Japan, 2009. IEEE.

[Pellegrini *et al.*, 2009b] 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*, Piscataway, NJ, 2009. IEEE Computer Society.

[Quan *et al.*, 2021] Ruijie Quan, Linchao Zhu, Yu Wu, and Yi Yang. Holistic lstm for pedestrian trajectory prediction. *IEEE Transactions on Image Processing*, 30:3229–3239, 2021.

[Robicquet *et al.*, 2016] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European Conference on Computer Vision*, Amsterdam, The Netherlands, 2016. Berlin: Springer.

[Saadatnejad *et al.*, 2024] Saeed Saadatnejad, Yang Gao, Kaouther Messaoud, and Alexandre Alahi. Social-transmotion: Promptable human trajectory prediction. In *The Twelfth International Conference on Learning Representations*, Vienna, Austria, 2024. Ithaca.

[Sadeghian *et al.*, 2019] 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1349–1358, Long Beach,CA, USA, 2019. Piscataway, NJ: IEEE.

[Shi *et al.*, 2021] Liushuai Shi, Le Wang, Chengjiang Long, Sanping Zhou, Mo Zhou, Zhenxing Niu, and Gang Hua. Sgcn:sparse graph convolution network for pedestrian trajectory prediction. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8990–8999, Nashville, TN, USA, 2021. IEEE.

[Shilun *et al.*, 2021] Li Shilun, Cai Tracy, and Li Jiayi. Trajectory prediction using generative adversarial network in multi-class scenarios. *CoRR*, abs/2110.11401(57), 2021.

[Tao *et al.*, 2020] Chaofan Tao, Qinhong Jiang, Lixin Duan, and Ping Luo. Dynamic and static context-aware lstm for multi-agent motion prediction. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI*, page 547–563, Berlin, Heidelberg, 2020. Springer-Verlag.

[Treue, 2003] Stefan Treue. Visual attention: the where, what, how and why of saliency. *Current Opinion in Neurobiology*, 13(4):428–432, 2003.

[Wang *et al.*, 2022] Jingke Wang, Tengju Ye, Ziqing Gu, and Junbo Chen. Ltp: Lane-based trajectory prediction for autonomous driving. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17113–17121, New Orleans, USA, 2022. IEEE.

[Wong *et al.*, 2024a] Conghao Wong, Beihao Xia, Ziqian Zou, Yulong Wang, and Xinge You. Socialcircle: Learning the angle-based social interaction representation for pedestrian trajectory prediction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19005–19015, Seattle, WA, USA, 2024. IEEE.

[Wong *et al.*, 2024b] Conghao Wong, Beihao Xia, Ziqian Zou, and Xinge You. Socialcircle+: Learning the angle-based conditioned interaction representation for pedestrian trajectory prediction, 2024.

[Xia *et al.*, 2022] Beihao Xia, Conghao Wong, Qinmu Peng, Wei Yuan, and Xinge You. Cscnet: Contextual semantic consistency network for trajectory prediction in crowded spaces. *Pattern Recognition*, 126:108552, 2022.

[Xu *et al.*, 2024] Baowen Xu, Xuelei Wang, Shuo Li, Jingwei Li, and Chengbao Liu. Social-cvae: Pedestrian trajectory prediction using conditional variational auto-encoder. In *Neural Information Processing*, volume 1962 of *Communications in Computer and Information Scienc*, page 476–489, Singapore, 2024. Springer Nature Singapore.

[Zhang *et al.*, 2023] Kunpeng Zhang, Liang Zhao, Chengxiang Dong, Lan Wu, and Liang Zheng. Ai-tp: Attention-based interaction-aware trajectory prediction for autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 8(1):73–83, 2023.