Physica A 663 (2025) 130435

Contents lists available at ScienceDirect

Physica A: Statistical Mechanics and its Applications

journal homepage: www.elsevier.com/locate/physa

SocialTrans: Transformer based social intentions interaction for pedestrian trajectory prediction

Kai Chen, Xiaodong Zhao *, Yujie Huang, Guoyu Fang

College of Mechanical and Electrical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China

ARTICLE INFO

Keywords:
Pedestrian trajectory prediction
Social state interaction
Intention extraction
Optimizer

ABSTRACT

The prediction of pedestrian trajectories plays a crucial role in practical traffic scenarios. However, current methodologies have shortcomings, such as overlooking pedestrians' perception of motion information from neighbor groups, employing simplistic and fixed social state interaction models, and lacking in final position correction. To address these issues, SocialTrans is proposed. It utilizes global observations to model the motion states of pedestrians and their neighbors, constructing separate state tensors to encapsulate social interaction information between them. This design includes a Subject Intention Extraction Module and a Neighbor Perception Intentions Extraction Module, which operate in parallel throughout the observation period to facilitate deep interaction of social states rather than simple end-to-end external fusion. Furthermore, a trajectory prediction optimizer is developed to correct final position predictions and simulate pedestrian motion diversity through trajectory clustering. Experimental validation is conducted on the ETH/UCY and SDD public datasets to evaluate the effectiveness of the proposed approach. The results demonstrate the method's capability to learn historical trajectory information, achieve high-precision predictions, and achieve state-of-the-art performance, particularly outperforming existing SOTA models on the SDD dataset. The algorithm will be made available at https://github.com/XiaodZhao/SocialTrans.

1. Introduction

Pedestrian trajectory prediction (PTP) involves forecasting potential movement paths of pedestrians within a future time frame based on historical positional data. This technology has broad applications, including autonomous driving and human-machine interaction in smart manufacturing. With continued advancements in deep learning models and their integration into intelligent systems, the development of effective PTP models has become increasingly feasible. Analyzing pedestrians' historical trajectory states through deep learning and incorporating scenario-based assessments of short-term intentions [1–5] have shown promise in improving the accuracy of PTP models, thus enhancing traffic safety. This capability also facilitates safety alerts in smart factories, helping prevent hazardous incidents. As a result, achieving high accuracy in PTP remains a significant focus of ongoing research.

Human behavior exhibits a significant degree of independence and adaptability, particularly in social environments, where pedestrians' movement paths are influenced by numerous intricate factors. While in motion, individuals are not only influenced by their immediate surroundings and objects but also may anticipate the actions of nearby individuals, prompting adjustments to their planned routes. Furthermore, when faced with unforeseen circumstances, they may rely on instinctive movement decisions. Additionally, even

* Corresponding author.
E-mail address: xdzhao@nuaa.edu.cn (X. Zhao).

https://doi.org/10.1016/j.physa.2025.130435
Received 12 December 2024; Received in revised form 7 February 2025;
Available online 12 February 2025

within the same environment, individuals may pursue different trajectories due to varying intentions; thus, complicating PTP due to this unpredictability. Therefore, as shown in Fig. 1, it is crucial to focus on the interaction of the pedestrians' intention information.

PTP work can be classified into two categories: deep learning models and knowledge-driven models [6]. Early approaches [7–20] encountered challenges in capturing complex environmental and latent state information. With advancements in hardware technology, these methods have gradually been replaced by data-driven deep learning techniques. Current methods have addressed various challenges. Long short-term memory (LSTM) based models [5,21–29] analyze individual movement states and social interaction behavior among subjects using specialized memory units. However, these models treat all interaction information equally, which does not align with social behavior norms. Additionally, they only consider memory from the previous moment, overlooking potential key information from future instances. Attention mechanism based models [30–36] have been refined to address imbalances in intentions information between subjects and neighbors. However, these improvements are limited to external interactions within attention mechanisms and fail to fully capture the intrinsic connections between subjects and their neighbors. Considering that it forces the computation of similarity between all query-keyword pairs, more computational resources are wasted on the processing of very weakly important long-distance pedestrian information. Inspired by the uncertainty of pedestrian motion, Variational autoencoder (VAE) based models [37–42] use latent variables generated over time for trajectory distribution estimation. Although they have resulted in improvements in the accuracy of prediction, the social analysis of neighbor relationships remains insufficiently thorough, and features processing overlooks potential future factors. Furthermore, existing methods encounter a common challenge where errors gradually increase with the lengthening of the prediction period, regardless of prediction effectiveness.

In response to the challenges identified in the aforementioned methods, this paper presents four key contributions:

(1) In addressing the fluctuating number of pedestrians within the scenario, we adopt motion state decomposition for modeling both the motion states of a selected individual and those of neighbor groups separately. We construct tensors for both the individual's state and the states of neighbor groups, adjusting them according to the varying pedestrian count in the scenario. Furthermore, latent perceptual information is incorporated into these tensors, providing the basis for intention extraction;
(2) The SocialTrans network is developed, incorporating the Subject Intention Extraction (SIE) Module and the Neighbor Perception Intentions Extraction (NPIE) Module. These modules operate simultaneously across time steps, fostering profound interaction of social states within them and conducting intention extraction. This facilitates more effective and precise learning of interaction information between the subject and neighbors. Ultimately, the interaction result decoder processes this information to derive trajectory prediction outcomes;
(3) We developed a trajectory prediction optimizer that considers both the distance error between each moment and the actual trajectory and the angle information between the distance vector determined by adjacent time points and the real trajectory. This addresses the challenge of error correlation with the prediction period. Furthermore, we apply cluster analysis to the multiple trajectory prediction results to replicate the diversity of pedestrian movement in the scenario;
(4) We assessed SocialTrans and contrasted it with existing methods using the openly accessible datasets ETH/UCY. The experimental outcomes indicate that SocialTrans achieves state-of-the-art performance, particularly in ETH scenarios, showcasing an average reduction in prediction error of about 40 % compared to existing methods. Moreover, when tested on the SDD dataset, it demonstrated an average reduction in prediction error of around 50 % compared to existing methods.

2. Related works

When considering object relations, PTP research can be split into two models: person-person models and person-space models.

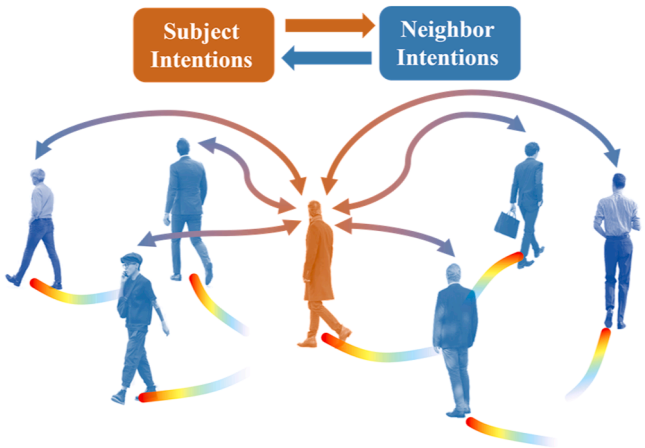

Fig. 1. Illustration of intentions interacion. It is acknowledgeed that intention information is crucial for the trajectory prediction task. We propose SocialTrans, which analyses historical trajectories, extracts pedestrian social intentions information and deeply interacts with them, finally outputs pedestrian trajectories that conform to social norms.

While the latter analyzes pedestrian trajectories within particular scenarios [43–45], relying on environmental cues, we assert that understanding the highly autonomous interaction of human movement is pivotal for trajectory prediction.

Initially, conventional person—person interaction models are utilized to tackle trajectory prediction challenges. They are mainly knowledge-driven and can be broadly categorized into models based on social force, geometric analysis, Markov estimation and game theory. **①Social force based models.** Helbing et al. [7] analyses pedestrian trajectories inspired by the attractive and repulsive forces of microscopic particles. However, it is sensitive to changes in model parameters and exhibits low generalizability; **②Geometric analysis based models.** PORCA[8] and ORCA[9] examine the geometric structures of observed objects, transforming trajectory analysis of person-person interactions into optimization problems; **③Markov estimation based models.**

The implicit Markov model for trajectory prediction proposed by Morris et al. [10] is employed for spatiotemporal probability modeling of diverse pedestrian trajectories. However, it heavily relies on time series and maintains stable state transition probabilities, which does not align with the complex variability of social norms. Consequently, this method is only suitable for short-term PTP; **④Game theory based models.** simulate interactions between pedestrian flows [11,12] and evacuation processes [13–15]. Some methods also consider latent factors like pedestrian attributes [16,17], walking groups [18,19], and stationary groups [20]. Essentially, most of these approaches are based on manual design and rules, which are not only unable to face subtle human interaction situations, but also make it difficult to perceive potential changes in the future.

In recent years, remarkable advancements have been achieved with learning-based data-driven models, thanks to the innovative progress in artificial intelligence technology. It is broadly categorized into LSTM, Attention mechanism and VAE based models.

2.1. LSTM based models

SocialLSTM [22] introduce a “social” pooling layer to enable sequences in close spatial proximity to share states with each other. However, it only considers information about pedestrian intentions within the grid, posing challenges in modeling complex temporal dependencies and simulating social interactions among all pedestrians. SocialGAN [25] addresses this by treating pedestrians in neighbor regions equally using pooling mechanisms. Nevertheless, this approach contradicts real-world pedestrian social norms, as subjects typically allocate varying levels of attention to surrounding neighbors.

2.2. Attention mechanism based models

STAR [31] significantly enhances temporal modeling by combining spatio-temporal perspective with graph convolution to address this issue. However, merely fusing spatio-temporal information in an end-to-end serial and parallel form does not allow for accurate

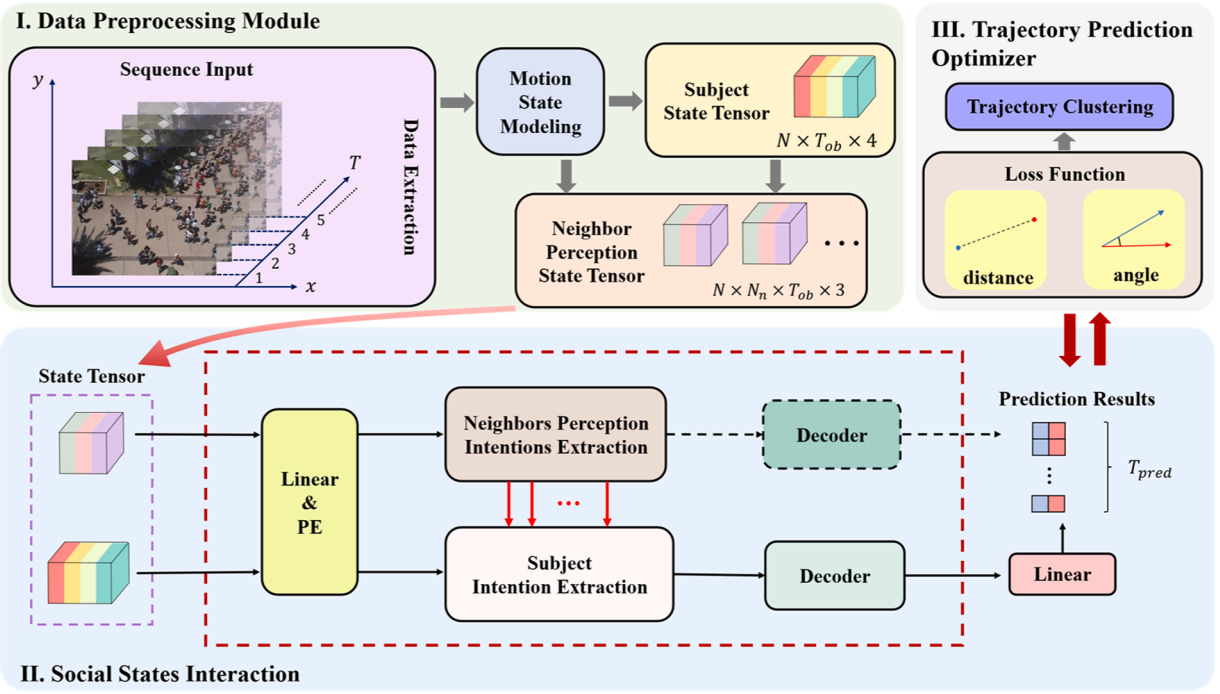

Fig. 2. Overall architecture of SocialTrans. In Section I, given the frame sequence of the pedestrian to be observed, multichannel state tensors for both the subject and its neighbors are constructed through data extraction and motion state modeling. These tensors are then input into the network in Section II, where the internal interaction of social states and intentions extraction occur. Finally, in Section III, the trajectory prediction optimizer refines the obtained trajectory results.

learning of high-quality social information. [34,35] performed the PTP task by analysing the social intentions of pedestrians through attention mechanism. However, they only process the intention of target pedestrians and neighbors with a single-attention module, which causes the model to learn ambiguous information about the intention, which in turn affects the final prediction results. Additionally, considerable effort is required to handle weakly influencing neighbors.

2.3. VAE based models

SocialVAE [1] employs an uncertainty model with RNN [46] as the primary architecture for predicting pedestrian trajectories. However, its social features only consider the information of the current moment, neglecting potential future features of neighbors that could impact pedestrian trajectories, posing challenges for modeling high-precision time series. Furthermore, these methods encounter a common challenge where errors are positively correlated with the prediction period, regardless of prediction effectiveness.

To address the above problems in knowledge-driven and data-driven models. We carry out research in three aspects: data scenario modelling, network design and optimization of prediction results. Finally, SocialTrans: Transformer Based Social Intentions Interaction for Pedestrian Trajectory Prediction is proposed.

3. Proposed approach

In PTP, capturing complex temporal dependencies over time is of utmost importance. Current models often focus solely on pedestrian trajectories at the present moment, overlooking potential crucial information in future instances. In dynamic environments with a continuously changing number of neighbors, preserving social information and comprehensively learning intricate interactions among neighbors is vital. Furthermore, in scenarios where pedestrians move toward each other, improving the accuracy of final position predictions is critical for enabling pedestrians to make timely decisions. These considerations motivate our proposal of the SocialTrans to tackle these challenges.

Our approach utilizes a transformer network to forecast the future trajectory distribution of each pedestrian, using the provided global historical observations. As depicted in Fig. 2, it comprises three primary modules: the data preprocessing module, the social state interaction module, and the trajectory prediction optimizer. In a scenario comprising (N_n+1) pedestrians, let $\{p_i^t\}_{t=+1}^{T_{ob}}$ represent the trajectory sequence of subject i during the observation period T_{ob}. Here, p_i^t denotes the two-dimensional spatial coordinates of subject i at time step t. By globally analyzing the trajectory information of all pedestrians during the period T_{ob}, we can construct state tensors for both the subject and its neighbor groups, incorporating potential future information. Using our custom-designed network structure, the trajectory sequence $\{p_i^t\}_{t=T_{ob}+1}^{T_{pred}}$ can be predicted for subject i. Angle information is backpropagated during training to refine the predicted trajectories

3.1. Data preprocessing module

In real-world scenarios, pedestrians encounter varying numbers of neighbors, a dynamic aspect often overlooked by methods such as those outlined in [22,25], which rely on pooling mechanisms. These mechanisms can lead to information loss and struggle to handle the complexities of changing neighbor counts. Hence, accurately modeling different neighbor numbers is essential for precise trajectory prediction. Inspired by [21], we introduce the data extraction module and a motion state module. By inputting frames of the pedestrian to be predicted, we extract data and analyze pedestrian motion in the scenario. This process constructs a three-dimensional state tensor for the target subject and a four-dimensional state tensor for their neighbors, capturing trajectory changes and interaction information from both perspectives.

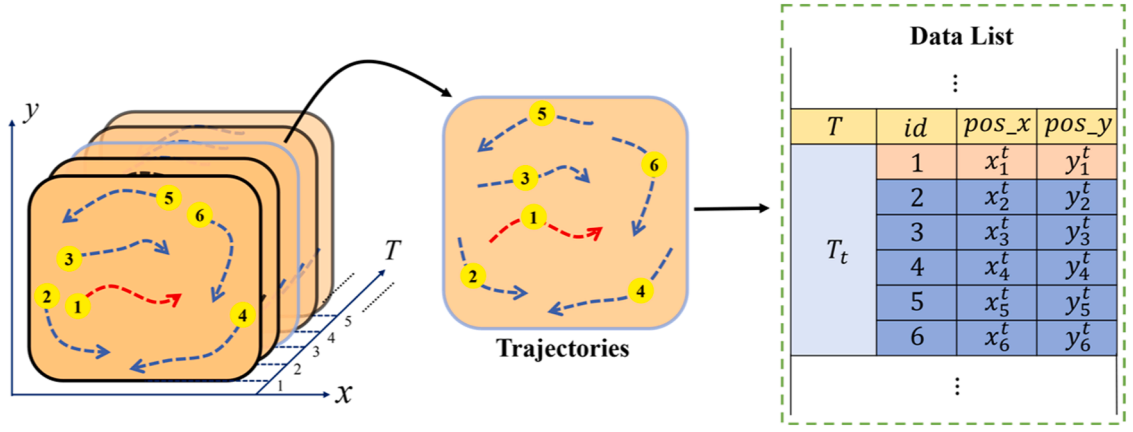

Fig. 3. Schematic diagram of the data extraction module. In order to cope with the changing number of pedestrians in the scenario, the current positions of all pedestrians in each frame are extracted under global observation to construct the data list.

Data Extraction. Redundant information and noise in the original input images may have unpredictable challenges for the PTP. Therefore, focusing on pedestrians within the images and retaining complete information for each of them is vital. For each observed frame, we directly extract the current positions of all pedestrians to form a data list. Fig. 3 exemplifies the scenario at T = 3, with pedestrian 1 selected as the subject for observation. We analyze one target subject at a time, while temporarily treating other pedestrians as global neighbors. Extracting position information of various pedestrians from the input image sequence establishes the data foundation for analyzing pedestrian motion states and constructing state tensors.

Motion State Module. Given that pedestrian trajectories are influenced by their perception of surrounding neighbors and neighbor movements, we introduce pedestrian perception information and develop a motion state module. This module captures the relationships between selected individuals and their neighbor groups. As depicted in Fig. 4, the motion states of the selected subject i and its neighbor j are decomposed, integrating perception information to construct the state tensor of the individual and its neighbor groups further.

Subject Motion State. The position of subject i in the scenario at moment t is labeled as (x_i^t, y_i^t), while tangential velocity is denoted as v_{ix}^t, and normal velocity denoted as v_{iy}^t. According to the processing in [43], the current position (x_i^t, y_i^t), is necessary, but the efficient and concise extraction of the subject's motion state is a primary concern. Although velocity and acceleration are identified as crucial in [55], directly utilizing individual pedestrians' forward direction, velocity, and acceleration as motion state inputs, without preprocessing as in SocialVAE [1], can lead to less focused network training. Given our relatively short observation time period, the velocity state of the pedestrian over a brief duration holds greater significance for trajectory prediction. Additionally, the computational costs incurred by integrating acceleration information outweigh the benefits throughout the entire process [31]. Therefore, we opt to describe the motion state based on the subject's position and velocity vectors, as depicted in (1) and (2). Consequently, the motion state of the currently selected subject i at moment t is denoted as $(x_i^t, y_i^t, v_i^t, \theta_i^t)$, then the state tensor of the subject in the whole observation period T_{ob} is $\mathbf{S} \in \mathbb{R}^{N \times T_{ob} \times 4}$, where N denotes the batch of subjects to be processed.

$$v_i^t = \sqrt{\left(v_{ix}^t\right)^2 + \left(v_{iy}^t\right)^2} \tag{1}$$

$$\theta_i^t = \arctan\left(\frac{v_{iy}^t}{v_{ix}^t}\right) \tag{2}$$

Neighbor Perception Motion State. After constructing the state tensor for subject i, we follow a similar procedure to decompose the motion state of each individual within the neighbor groups. Subsequently, we aim to articulate the interaction dynamics between the subject and the neighbor groups. As mentioned in the related work [31], the distance $\left\|\overrightarrow{l_{ij}^t}\right\|$ between subject i and neighbor j at moment t stands out as the crucial metric employed to depict their ongoing interaction [1], as depicted in (3). However, relying solely on distance fails to capture the full scope of social perception information. Moreover, when multiple paths exist between the subject and the interacting neighbor at the same distance, it complicates the characterization of social movement states. To enhance our understanding of social behaviors and more accurately represent the social dynamics, we address this challenge by incorporating the angle of their velocities θ_{ij}^t as illustrated in (4).

$$\overrightarrow{l_{ij}^t} = \left(x_j^t - x_i^t, y_j^t - y_i^t\right) \tag{3}$$

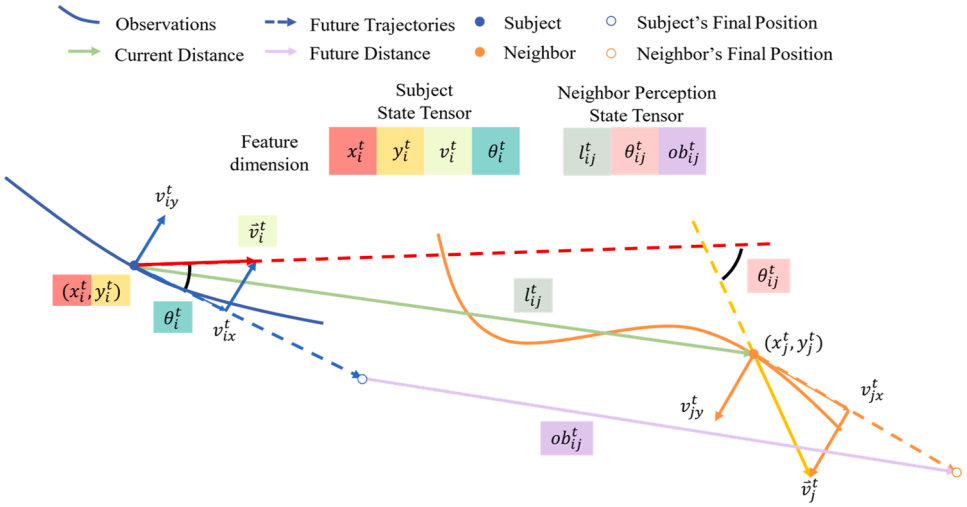

Fig. 4. Schematic diagram of trajectory analysis. Motion state analysis is performed for the selected subject i and neighbor j, while the trajectory changes are perceived at the current moment t for the final moment in the historical observation period.

$$\theta_{ij}^{t} = \arccos\left(\frac{\overrightarrow{v_i^t}\cdot\overrightarrow{v_j^t}}{\left\|\overrightarrow{v_i^t}\right\|\left\|\overrightarrow{v_j^t}\right\|}\right) \tag{4}$$

In contrast to numerous existing approaches [22,47,54], our method differs by integrating the social decisions of both the subject and its neighbors within the scenario. Unlike methods that rely solely on a combination of historical and current information, our approach also includes the anticipation of potential future impacts across the temporal domain. According to the law of inertia, an object in motion maintains its direction and speed. Therefore, we assume that pedestrians will continue to move along their tangential velocity unless they are influenced by unexpected factors during their travel. By examining the size of the time period (T_{ob}-t) at the current moment relative to the end of the observation period, as well as the time ξ required for subject i and neighbor j to meet at their current speed, we select the shorter duration for their continued movement along the tangential direction with their tangential velocity. This process, correlated with the distance at the current moment, yields the final distance ob_{ij}^t under the observation period as in (5), (6). The final motion-aware state of neighbor j at moment t is denoted as $(\left\|\overrightarrow{l_{ij}^t}\right\|, \theta_{ij}^t, ob_{ij}^t)$, then the state tensor of the neighbor groups in the whole observation period T_{ob} is $\mathbf{O} \in \mathbb{R}^{N\times N_n\times T_{ob}\times 3}$, where N_n indicates the number of neighbors.

$$\xi = \left|\frac{\overrightarrow{l_{ij}^t}\cdot\overrightarrow{v_{ij}^t}}{\left\|\overrightarrow{v_{ij}^t}\right\|^2}\right| \tag{5}$$

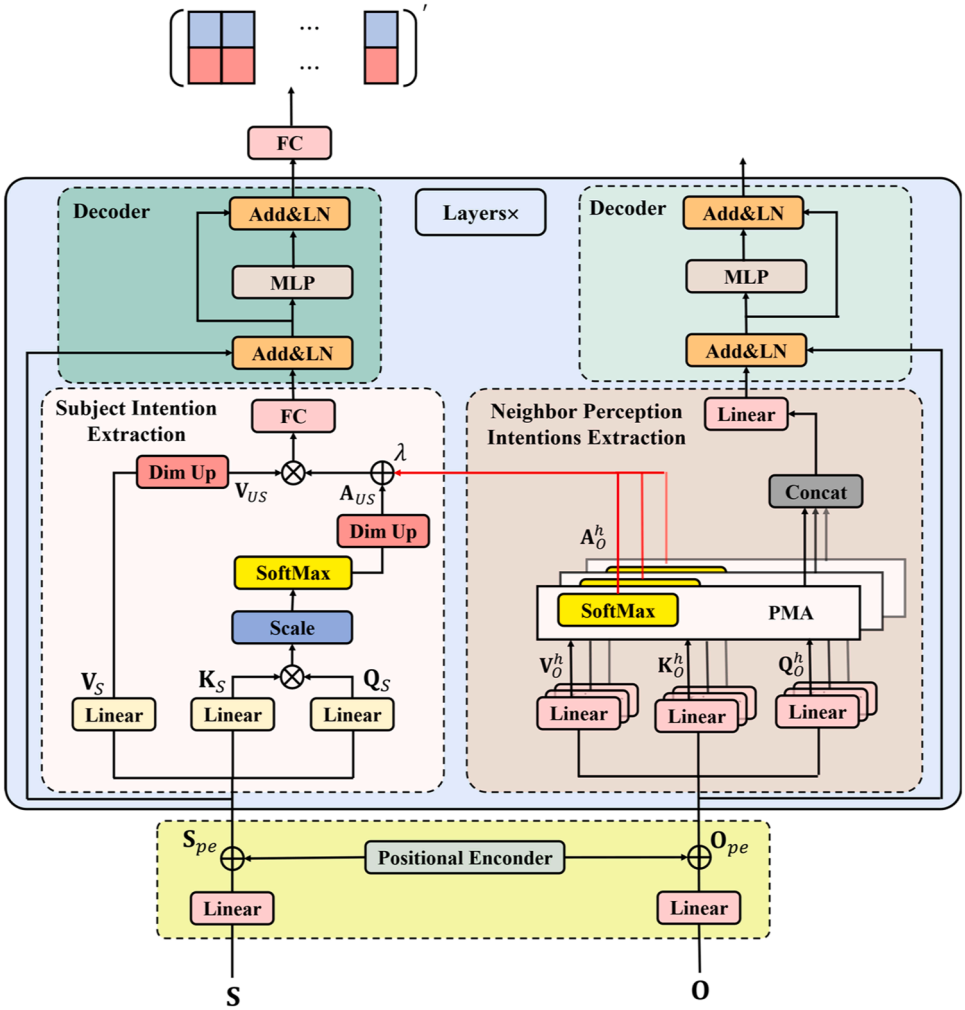

Fig. 5. Network structure diagram. The state tensors of the subject and neighbor groups are inputted into the subject intention extraction (SIE) and neighbor perception intentions extraction (NPIE) modules through a linear layer and position encoding, respectively, and then decoded by an interaction result decoder to obtain preliminary prediction results.

$$ob_{ij}^{t} = \left\| \overrightarrow{l_{ij}^{t}} + \min(\xi, T_{ob} - t) \cdot \left(\overrightarrow{v_{jx}^{t}} - \overrightarrow{v_{ix}^{t}} \right) \right\| \tag{6}$$

We created a data list by extracting information from the input image sequences and developed a motion state module to analyze pedestrian motion in the scenario and integrate perceptual information. Subsequently, we constructed the state tensor **S** of the subject and the state tensor **O** of the neighbor groups within T_{ob}. For scenarios in the historical observation period where there are no neighbors or only a single subject, we handle them as follows: we introduce an invalid neighbor as a placeholder, with motion feature values (x, y, v_x, v_y) set to 1e9. By assigning this extreme value (1e9), which exceeds the physical range, we create distinguishable invalid sample identifiers. In Section 3.2, the proposed Perception Mask Attention module will filter out these marked invalid inputs. It should be noted that the above construction process is all automated, without manual labelling. This not only improves the efficiency of data processing, but also establishes a good foundation for the subsequent extraction and processing of social interaction states.

3.2. Social states interaction

In order to clearly extract the intentions information of pedestrians and realise the deep interaction of social information, we design the social states interaction network structure as shown in Fig. 5. In order to maintain the temporal information, we process the state tensor **S** of the subject and the state tensor **O** of the neighbor groups by linear transformation and positional encoding. Previous studies [25,29,39] have typically applied the attention mechanism solely at the last frame of an observation sequence. Social VAE [1] only uses the attention mechanism at each moment separately.

The repetitive superposition of the attention mechanism still does not take into account the connection between each moment well. Therefore, in order to optimize both the navigation strategies of subject pedestrians and the social influence of their neighbors from a macroscopic point of view, we designed two intention extraction networks that operate concurrently throughout the observation period. Furthermore, unlike in [31], where processing results are simply connected in series or parallel, our approach involves overlaying the weights of the SIE onto those of the NPIE, facilitating internal interaction of social interaction information throughout the entire processing pipeline.

Neighbor Perception Intentions Extraction (NPIE). The input of NPIE $\mathbf{O}_{pe} \in \mathbb{R}^{N \times N_n \times T_{ob} \times d_{mdoel}}$ is derived through a linear transformation and positional encoding following **O**. The subject appears in every frame of the historical observation period. For other agents in each frame, they are stored in a neighbor list. If a certain neighbor does not appear in the current historical frame, its motion feature information (x, y, v_x, v_y) is set to 1e9, in order to match the feature dimensions of the state tensor **O**. We project to query vector $\mathbf{Q}_O^h$, key vector $\mathbf{K}_O^h$ and value vector $\mathbf{V}_O^h$ [31] with a few different, learned linear projections ε times. $\mathbf{Q}_O^h, \quad \mathbf{K}_O^h, \mathbf{V}_O^h \in \mathbb{R}^{N \times N_n \times 1 \times T_{ob} \times d_\varepsilon}$, $h = 1, \ldots, \varepsilon$, and d_ε is the feature dimension after ε projections. According to (7), we obtain the attention weight $\mathbf{A}_O^h$ for each head in the neighbor cluster. By examining the relationship between the distance $\left\| \overrightarrow{l_{ij}^{t}} \right\|$ and the observation radius r_i between subject i and each global neighbor j throughout the entire T_{ob} period, we designate distances larger than r_i as non-neighbors, assigning them a value of 0 in the perception mask. Distances within the observation radius are treated as neighbors and assigned a value of 1, resulting in a sparse perception mask. As shown in Fig. 6, it sparsifies the huge similarity information of $\mathbf{K}_O^h$ and $\mathbf{V}_O^h$, eliminates the computational overhead of weakly important distant pedestrians during processing, reduces the amount of computation and memory usage, and thus reduces the computational complexity. Considering that when a certain neighbor and a subject are far enough away from each other, we comprehensively consider shedding the impact of the neighbor on the subject's short-term trajectory to make the network computation more efficient and convenient.

We get the intention matrix $\mathbf{H} \in \mathbb{R}^{N \times N_n \times T_{ob} \times d_{mdoel}}$ of neighbor groups in (7), (8) and (9). To realize social state interaction, $\{\mathbf{A}_O^h\}_{h=1}^{\varepsilon}$ is

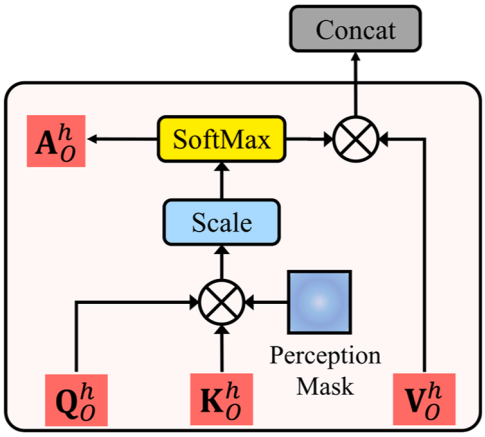

Fig. 6. Perception mask attention mechanism. The sparse perceptual mask is applied to each head to obtain the attention weight $\mathbf{A}_O^h$, highlighting key information while minimizing the computational overhead caused by weak influences.

used as one of the inputs to the SIE.

$$\mathbf{A}_O^h = \mathrm{Softmax}\left(\mathbf{M}_O \cdot \frac{\mathbf{K}_O^h \cdot \mathbf{Q}_O^h}{\sqrt{d_\varepsilon}}\right) \tag{7}$$

$$\mathrm{Att}_O^h = \mathbf{A}_O^h \cdot \mathbf{V}_O^h \tag{8}$$

$$\mathbf{H} = \left(\mathrm{Att}_O^1 \circ \mathrm{Att}_O^2 \circ \cdots \circ \mathrm{Att}_O^h\right) \cdot \mathbf{W}_O \tag{9}$$

where $\mathbf{A}_O^h \in \mathbb{R}^{N\times N_n\times 1\times T_{ob}\times T_{ob}}$ denotes attention weight of the head h. $\mathbf{M}_O^h \in \mathbb{R}^{N\times N_n\times 1\times T_{ob}\times T_{ob}}$ is perception mask. $\mathrm{Att}_O^h \in \mathbb{R}^{N\times N_n\times 1\times T_{ob}\times T_h}$ is attention score of the head h, and $\mathbf{W}_O \in \mathbb{R}^{\varepsilon d_\varepsilon \times d_{mdoel}}$ is projection matrix. ' $\circ$ ' is used to concatenate the results of the attention functions of the ε heads indexed by Att_O^h.

Subject Intention Extraction (SIE). The input of SIE $\mathbf{S}_{pe} \in \mathbb{R}^{N\times T_{ob}\times d_{mdoel}}$ is derived through a linear transformation and positional encoding following $\mathbf{S}$. After applying three linear transformations, we obtain the query vector $\mathbf{Q}_S$, key vector $\mathbf{K}_S$, and value vector $\mathbf{V}_S$, where $\mathbf{Q}_S, \mathbf{K}_S, \mathbf{V}_S \in \mathbb{R}^{N\times T_{ob}\times d_{mdoel}}$. To eliminate the influence of random errors and outliers in the neighbor groups, we filter the number of neighbors N_n of $\{\mathbf{A}_O^h\}_{h=1}^{\varepsilon}$ and obtain filter matrix sequence $\{\overline{\mathbf{A}_O^h}\}_{h=1}^{\varepsilon}$, where $\overline{\mathbf{A}_O^h} \in \mathbb{R}^{N\times\varepsilon\times T_{ob}\times T_{ob}}$. Considering the match with $\overline{\mathbf{A}_O^h}$, we dimensionally expand $\mathbf{V}_S$ to obtain ascending dimension value vector $\mathbf{V}US \in \mathbb{R}^{N\times\varepsilon\times T_{ob}\times d_{mdoel}}$. Meanwhile, as shown in (10), after obtaining the attention weight $\mathbf{A}_S$ in SIE, the same dimensional expansion is applied to it to obtain ascending dimension attention weight $\mathbf{A}US \in \mathbb{R}^{N\times\varepsilon\times T_{ob}\times T_{ob}}$. We introduce the scaling factor λ (this parameter is ultimately obtained through training) to regulate the effect of the introduced neighbor groups attention weights and obtain the final attention score $\mathbf{F}$ through full connectivity, as shown in (11).

$$\mathbf{A}_S = \mathrm{Softmax}\left(\frac{\mathbf{K}_S \cdot \mathbf{Q}_S}{\sqrt{d_{model}}}\right) \tag{10}$$

$$\mathbf{F} = \left(\mathbf{A}_{US} + \lambda\overline{\mathbf{A}_O^h}\right) \cdot \mathbf{V}_{US} \cdot \mathbf{W}_S \tag{11}$$

where $\mathbf{A}S \in \mathbb{R}^{N\times T_{ob}\times T_{ob}}$ is the attention weight in NIE and $\mathbf{W}S \in \mathbb{R}^{N\times T_{ob}\times d_{mdoel}}$ is the learned linear transformation matrix.

Interaction Results Decoder. To address the challenges associated with vanishing and exploding gradients and ensure smoother gradients, the initial segment of the encoder is structured as a residual connection. This section integrates a dropout function, an addition operation, and layer normalization. After processing the inputs, intention normalization matrix $\mathbf{C}_S$ of subject and $\mathbf{C}_O$ of neighbor groups are obtained respectively as in (12), (13).

$$\mathbf{C}_S = \mathrm{LN}\left(\mathrm{Dropout}(\mathbf{H}) + \mathbf{S}_{pe}\right) \tag{12}$$

$$\mathbf{C}_O = \mathrm{LN}\left(\mathrm{Dropout}(\mathbf{F}) + \mathbf{O}_{pe}\right) \tag{13}$$

The $\mathbf{C}_S$ and $\mathbf{C}_O$ are fed into the multilayer perceptron for processing as in (14), (15). It has been shown in [56,57] that multilayer perceptron of this structure help to improve the embedding quality.

$$\mathbf{C}_S' = \mathrm{FC}(\mathrm{ReLU}(\mathrm{FC}(\mathbf{C}_S))) \tag{14}$$

$$\mathbf{C}_O' = \mathrm{FC}(\mathrm{ReLU}(\mathrm{FC}(\mathbf{C}_O))) \tag{15}$$

where $\mathbf{C}_S'$ and $\mathbf{C}_O'$ are the processing results of $\mathbf{C}_S$ and $\mathbf{C}_O$ after the multilayer perceptron, respectively.

To slow the model degradation, we have a layer of residual links superimposed after the multilayer perceptron. The final decoding matrix $\mathbf{C}_S''$ of subject and $\mathbf{C}_O''$ of neighbor groups are obtained as in (16) and (17), respectively.

$$\mathbf{C}_S'' = \mathrm{LN}\left(\mathrm{Dropout}\left(\mathbf{C}_S' + \mathbf{C}_S\right)\right) \tag{16}$$

$$\mathbf{C}_O'' = \mathrm{LN}\left(\mathrm{Dropout}\left(\mathbf{C}_O' + \mathbf{C}_O\right)\right) \tag{17}$$

In the final layer of the decoder, $\mathbf{C}_S''$ from the branch containing $\mathbf{S}$ is passed through a fully connected process to obtain the trajectory representation matrix $\mathbf{T} \in \mathbb{R}^{l\times T_{pred}\times 2}$, where l denotes the number of predicted trajectories and T_{pred} represents the number of time steps for the predicted trajectories.

In this approach, the network utilizes the state tensor $\mathbf{S}$ of the subject, and the state tensor $\mathbf{O}$ of the neighbor groups. These tensors undergo intentions extraction and deep internal interactions of social information within the NPIE and SIE modules. Subsequently, the decoder decodes the processed output, resulting in trajectory predictions. These predicted trajectories that closely align with ground truth trajectories and adhere to social norms.

3.3. Trajectory prediction optimizer

While prior work [1,31,47] primarily focuses on predicting entire trajectories, it overlooks the tendency for prediction errors to escalate with longer prediction periods. Additionally, it neglects the significance of accurately predicting final positions for optimizing the rapid decision-making of intelligent systems, especially in scenarios involving opposing directions without meeting. By directly predicting final positions across multiple periods with lower information value, we circumvent the need for extensive computations, a critical consideration for resource-intensive intelligences.

The directional dynamics of pedestrians walking in opposite versus the same directions have distinct effects on their motion states. Hence, during the backpropagation of predicted trajectories, we posit that, in addition to position error, trajectory direction error significantly influences prediction accuracy. Thus, we introduce angular loss, derived from existing trajectory distance error [21], to supervise the training process during backpropagation. This approach facilitates trajectory

correction and yields improved prediction outcomes within the T_{pred}. As illustrated in Fig. 7, the red trajectory denotes the actual trajectory, the blue one signifies the predicted trajectory, and we select the (T_{ob}+t) moment to assess both distance and angle errors.

For the distance, we select the predicted position of each period for comparison with the true position. The Euclidean distance is calculated, and thus, the distance prediction error for each period is obtained as expressed in (18).

$$L_{dis} = \sum_{t=T_{ob}+1}^{T_{pred}} \left\| p_i^t - p_{igt}^t \right\| \tag{18}$$

where p_{igt}^t is the true position of subject i at moment t

To address angular considerations, we introduce the final moment of the observation period, T_{ob}. The direction vector is formed by connecting each predicted moment to the preceding one. Subsequently, we evaluate the angle relative to the direction vector of the true trajectory, as described in Eq. (19).

For the treatment of angles, we introduce the last moment of the observation period. The direction vector is constructed by linking each predicted moment to its preceding moment, and then the angle with the direction vector of the true trajectory is examined as expressed in (20).

$$L_{angle} = \sum_{t=T_{ob}}^{T_{pred}-1} \arccos\left(\frac{\left(p_i^t - p_i^{t+1}\right) \cdot \left(p_{igt}^t - p_{igt}^{t+1}\right)}{\left\| p_i^t - p_i^{t+1} \right\| \left\| p_{igt}^t - p_{igt}^{t+1} \right\|} \right) \tag{19}$$

$$L = L_{dis} + L_{angle} \tag{20}$$

The model's objective loss function, as shown in (20), aims to minimize the loss function L by backpropagating both the distance error and angular error between each prediction and the corresponding ground truth to the network. In Section 4.5, ablation experiments will be conducted to demonstrate the optimization effect of this loss function.

For the final set of multiple trajectories, we apply DBSCAN (Density-Based Spatial Clustering of Applications with Noise), a density-based clustering algorithm. Unlike the final position clustering approach in [1], this method dynamically determines the number of categories, providing increased flexibility and better alignment with pedestrian autonomous navigation strategies by avoiding the need to artificially preset the number of categories.

4. Experiments

In this section, we investigate the entire implementation process of SocialTrans. We utilize datasets from ETH/UCY and SDD for experimentation, along with a detailed description of evaluation metrics. SocialTrans is compared with existing trajectory predictors, including state-of-the-art (SOTA) models [1]. Additionally, through ablation studies, we examine the specific impact of each component, providing a deeper understanding of key aspects of trajectory prediction tasks.

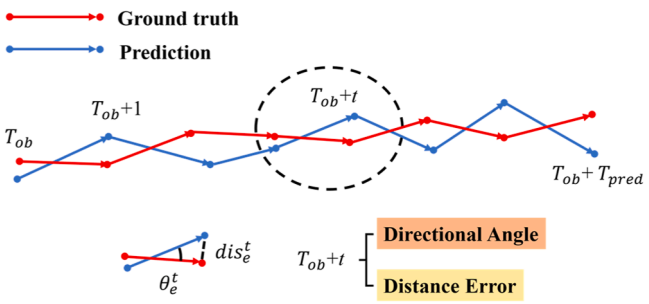

Fig. 7. Loss function for multidimensional information fusion. This function contains distance information and angle information to supervise and correct the prediction results of the network in backpropagation so that it can maintain good performance in the prediction of the final position.

4.1. Datasets and metrics

Datasets. The Stanford Drone Dataset (SDD) [52] contains the movement paths of 5232 individuals across eight different scenarios, making it one of the largest datasets available. It includes various entities like pedestrians, cyclists, skateboarders, cars, buses, and golf carts, navigating through real-world outdoor settings such as university campuses. On the other hand, the ETH/UCY benchmark [48, 49] documents the trajectories of 1536 pedestrians across five distinct scenarios: ETH, HOTEL, UNIV, ZARA1, and ZARA2. These datasets serve as valuable resources for trajectory data, capturing complex pedestrian dynamics like co-directional and counter-directional movements; thus, presenting significant challenges for trajectory prediction tasks.

Metrics. To assess our model's performance, we utilize the following two metrics:

(1) Average Displacement Error (ADE): This metric calculates the average Euclidean distance between the ground truth coordinates and the predicted coordinates across all time steps [50].

$$ADE = \frac{\sum_{i=1}^{num} \sum_{t=T_{ob}+1}^{T_{pred}} \left\| p_i^t - p_{igt}^t \right\|}{num \times T_{pred}} \tag{21}$$

Where *num* represents the number of pedestrians.

(2) Final Displacement Error (FDE): The Euclidean distance between the predicted points and the ground truth point at the final prediction time instant T_{pred} [22].

$$FDE = \frac{\sum_{i=1}^{num} \left\| p_i^{T_{pred}} - p_{igt}^{T_{pred}} \right\|}{num} \tag{22}$$

4.2. Implementation details

Our network architecture, training process, and prediction implementation rely on PyTorch. To mitigate the impact of pedestrians' inherent movement patterns, we employ data augmentation techniques such as horizontal flipping and rotation to augment the extracted data. Based on the processing of [58–60] we set the model parameters as follows: d_{model}= 512, ε= 8, d_{ε}= 64. λ is trained to obtain the optimal result with the minimum loss value. Considering previous work [1], the number of neighbors N_n is determined by the number of pedestrians present in the scene during the historical observation period. Since our network simultaneously performs the attentional mechanism for the entire period of historical observations, the number of frames for the observations and predictions is kept equal during training, where we set to T_{ob}=T_{pre} $_{\mathrm{d}}$= 8. Considering the diversity of pedestrian movements, some historical observations cannot correctly predict the future, so multiple trajectories are reasonable and conform to social norms. Therefore, we set the number of predicted trajectory bars *l* for the subject to 20 to provide rich choices. For each scene in the ETH/UCY and SDD datasets,

we defined specific training hyperparameters. In the ETH/UCY dataset, the initial learning rate for the Eth scene is set to 0.0005, with 600 training epochs and testing performed every 200 epochs. For the Hotel scene, the initial learning rate is 0.0007, with 600 training epochs and testing performed every 200 epochs. The Univ scene uses an initial learning rate of 0.0001, with 200 training epochs and testing performed every 100 epochs. For the Zara01 scene, the initial learning rate is 0.0003, with 600 training epochs and testing performed every 200 epochs. The Zara02 scene also has an initial learning rate of 0.0005, with 600 training epochs and testing performed every 200 epochs. In the case of the SDD dataset, the initial learning rate is 0.0007, with 600 training epochs and testing performed every 200 epochs.

Utilizing leave-one-out cross-validation, as described in previous research [28], the dataset mentioned in Section 4.1 is employed. The Adam optimizer [51], an extension of stochastic gradient descent, is utilized within the SocialTrans network for updating network weights during training. A learning rate of 0.001, dropout rate of 0.2, gradient clipping threshold of 10, and weight decay of 0.0001 are applied, with 1000 epochs performed per training session. During training, the best parameter model is identified based on the validation set, achieving the lowest ADE. The inference process involves observing 8 frames and predicting the subsequent 8 frames by utilizing the best parametric model. This process is executed on a machine equipped with two RTX3090 GPUs.

4.3. Quantitative results and analysis

In the SDD dataset, as described in [53], we perform trajectory prediction for each of the eight scenarios using image segmentation techniques to extract pedestrian positional information on a pixel-by-pixel basis. It is worth noting that, unlike previous experimental methods, we adopt the same approach of observing 8 frames and predicting 8 frames for other trajectory prediction models to ensure the fairness of the experiments. Our proposed SocialTrans method stands out, achieving a remarkable 52 % reduction in ADE and a 56 % decrease in FDE compared to the current state-of-the-art performance of 6.94/9.46.

As illustrated in Table 1, the analysis of the ETH/UCY datasets, especially in the HOTEL scenario marked by reduced crowding and simpler trajectory distributions, demonstrates the superior performance of the linear method over several LSTM-based deep learning approaches, including SocialGAN [25], SR-LSTM [27], SoPhie [28], and SocialWays [29]. However, large errors still occur in more complex scenarios with these methods. Temporal dependency issues inherent in

LSTM approaches are effectively addressed by Transformer-based methods, such as STAR [31] and TransformerTF [36], which demonstrate superior performance across both datasets. Nevertheless, these models often overlook the deeper social factors influencing pedestrian behavior, and their integration of spatio-temporal interactions remains relatively simplistic. VAE-based approaches like Trajectron+ + [38], SGNet-ED [39], BiTraP [40], and SocialVAE [1] (with or without FPC post-processing) have significantly reduced errors. While SocialVAE has achieved state-of-the-art (SOTA) results in many scenarios, its treatment of social features remains focused on the present moment, overlooking potential key future information. Its attention covers only isolated single moments, leading to weaker macro control of trajectory predictions and leaving room for improvement. Our SocialTrans compre-hensively addresses these issues, achieving SOTA performance on these datasets, particularly in the ETH scenario, with a 50 % improvement in ADE and a notable 72 % enhancement in the FDE.

4.4. Qualitative results and analysis

To preliminarily evaluate SocialTrans's predictive performance, we conducted trajectory visualization on the ETH, HOTEL, ZARA01, and ZARA02, comparing outcomes with those of SocialVAE. As shown in Fig. 8, the visualization results for the simple scenario are presented. In the HOTEL scenario, predictions from SocialVAE notably deviated from ground truth, especially in direction, while SocialTrans effectively inferred intent information from historical trajectories, resulting in accurate predictions. Specifically, in HOTEL(a) and HOTEL(d), where the subject exhibited multiple directional changes, SocialTrans accurately predicted trajectory alterations, closely aligning ground truth. This accuracy can be attributed to the effectiveness of our designed loss function. Regardless of neighbor density, SocialTrans effectively captured interaction cues among neighbors and produced precise predictions. Furthermore, in most subsequent phases, outcomes closely aligned with ground truth. Conversely, predictions by SocialVAE tended to cluster around the initial phase or produced short trajectories, indicating a failure to effectively grasp intent interaction information between the subject and its neighbors.

To better demonstrate the prediction effect of SocialTrans, we choose the ZARA02 and UNIV scenarios, which are rich in pedestrians and have a large number of pedestrians, for visualization, as shown in Fig. 9. Whether subjects remain stationary or maintain their initial motion state, SocialTrans effectively models complex pedestrian interactions and delivers more precise predictions compared to SocialVAE. In the ZARA02 scenario's first column, where both motorial and stationary motion states of neighbors are observed, SocialTrans captures subjects' intentions, yielding predictions that closely clustered like ground truth, while SocialVAE produces trajectories in a slow motorial state. In the UNIV scenario, characterized by a large number of neighbors, SocialTrans maintains accuracy even when some neighbor trajectories intersect with the subject. Notably, the subject appears to prioritize neighbors closer in contemporaneous distance, as evident from the transparency of gray trajectories. However, in the ZARA02 scenario's second column (d) and the UNIV scenario's first column (b), even neighbors at similar distances during the heterogeneous period do not receive higher attention scores. This highlights the effectiveness of our self-designed attention mechanism, which operates not only at individual moments but also globally throughout the entire observation period to comprehensively examine interactions.

To illustrate the functioning of our self-designed attention mechanism, attention maps are presented in Fig. 10 across three scenarios: the first row depicts HOTEL, the second row UNIV, and the third row ZARA. These maps visualize the attention weights of each

Table 1
Quantitative results of the correlation methods analyzed on the TWO datasets. All the work is realised using 8 frames of prediction for 8 frames and the average of the 20 predicted values for the best effect of ADE/FDE is selected. ^{D}: deterministic version of the model. ^{i}: requires image input. $^{\circ P}$: reproduces the results after optimizing the existing problem.

	SDD	ETH	HOTEL	UNIV	ZARA01	ZARA02
Linear	16.53/35.12	1.23/2.56	0.41/0.76	0.86/1.54	0.62/1.24	0.76/1.56
S-LSTM [22]	–	0.81/1.68	0.39/0.82	0.59/1.21	0.54/0.98	0.42/0.85
CIDNN [24]	–	1.17/2.07	1.12/1.69	0.60/1.19	0.84/0.94	0.45/0.92
S-GAN [25]	22.51/34.65	0.57/0.89	0.41/0.86	0.49/0.96	0.29/0.59	0.24/0.54
Trafficpredict [26]	–	5.14/8.85	2.32/3.35	3.07/5.97	3.46/6.17	3.35/6.89
SR-LSTM [27]	–	0.59/1.14	0.33/0.69	0.44/0.91	0.36/0.89	0.29/0.65
SoPhie^{i} [28]	12.31/23.62	0.64/1.27	0.69/1.56	0.49/1.16	0.26/0.57	0.34/0.65
Socialways [39]	–	0.37/0.56	0.35/0.59	0.51/1.27	0.39/0.57	0.45/0.88
MemoNet^{OP} [4]	7.24/9.54	0.37/0.54	0.12/0.20	0.21/0.42	0.18/0.24	0.12/0.19
STAR^{D}	–	0.52/1.06	0.23/0.45	0.38/0.84	0.28/0.66	0.40/0.81
STAR [31]	–	0.32/0.59	0.18/0.29	0.29/0.60	0.22/0.45	0.19/0.44
TransformerTF [36]	–	0.57/1.09	0.17/0.23	0.31/0.56	0.18/0.34	0.15/0.24
MANTRA [3]	7.83/13.68	0.44/0.91	0.17/0.28	0.33/0.68	0.19/0.35	0.16/0.24
PECNet [37]	8.09/13.76	0.51/0.82	0.16/0.24	0.32/0.52	0.20/0.34	0.14/0.26
Trajectron+ +^{OP} [38]	8.94/14.32	0.49/0.89	0.14/0.21	0.26/0.47	0.17/0.37	0.13/0.26
SGNet-ED^{OP} [39]	8.07/15.67	0.45/0.88	0.19/0.47	0.30/0.65	0.16/0.32	0.13/0.30
BiTraP^{OP} [40]	7.96/13.64	0.52/0.86	0.17/0.25	0.23/0.42	0.21/0.41	0.15/0.29
AgentFormer [41]	–	0.43/0.69	0.13/0.19	0.22/0.41	0.16/0.24	0.13/0.19
SocialVAE [1]	7.34/11.74	0.42/0.73	0.14/0.22	0.25/0.47	0.20/0.37	0.14/0.28
SocialVAE+FPC	6.94/9.46	0.35/0.59	0.13/0.19	0.21/0.36	0.17/0.29	0.13/0.22
SocialTrans	**3.30/4.11**	**0.22/0.18**	**0.11/0.12**	**0.17/0.16**	**0.16/0.18**	**0.10/0.11**

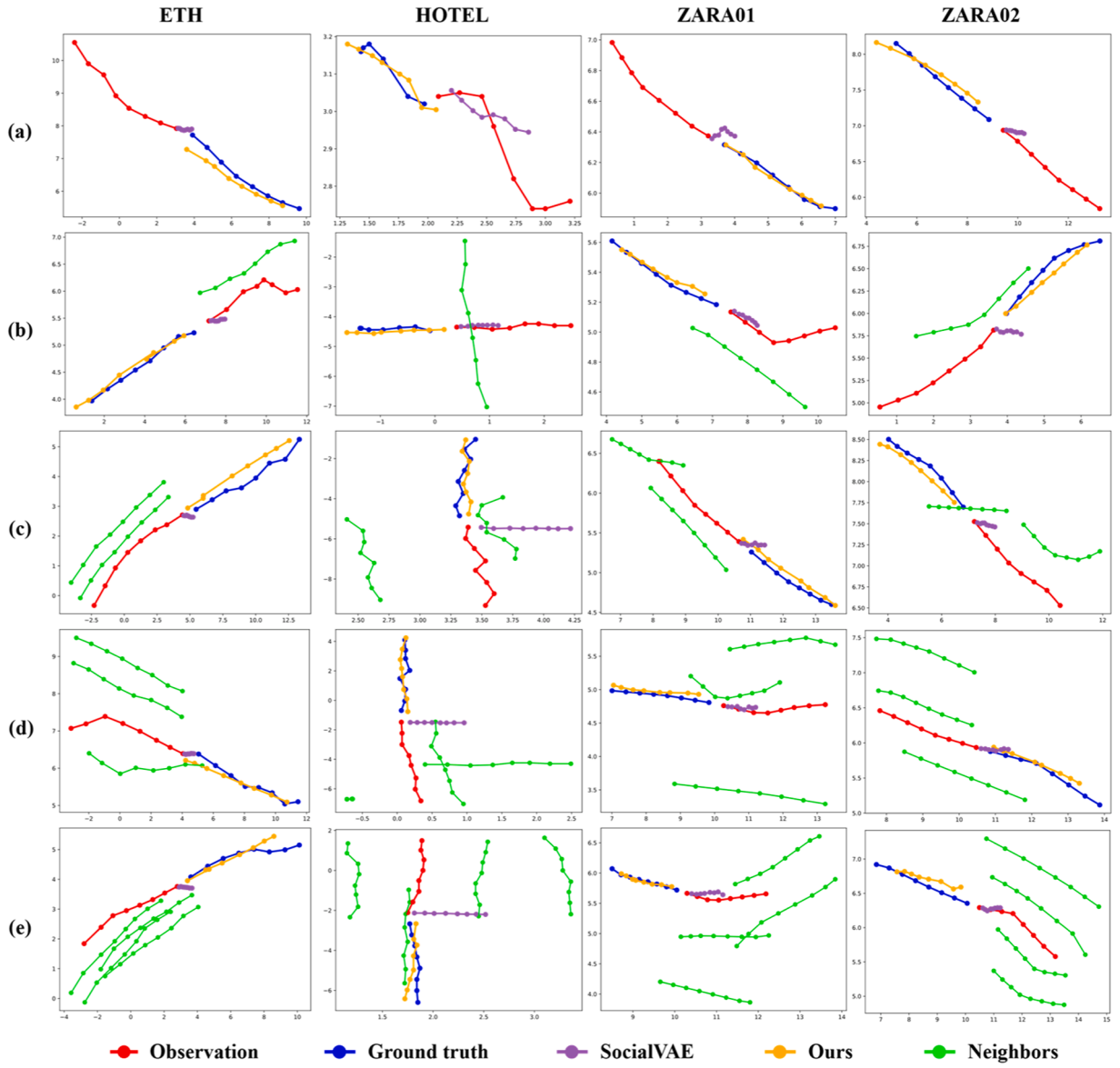

Fig. 8. Trajectories visualization in simple scenarios. We sequentially display predictions for the subject itself and neighbors increasing from one to four. SocialTrans outperforms the state-of-the-art (SOTA) model SocialVAE, in making more accurate predictions.

neighbor using gradient-filled circles, where the opacity and radius correlate with the neighbor's weight in their respective scenarios. Overall, subjects prioritize neighbors moving in the same direction and those approaching from opposite directions. Conversely, neighbors behind the subjects, regardless of whether they are moving toward or away from them post-encounter, receive minimal attention. Furthermore, influenced by DBSCAN, the number of decision trajectories for the subjects continuously adjusts in response to changes in scenarios and motion states.

For a more intuitive representation of the interaction between the subject and its neighbors, we present it visually in Fig. 11. In row (a), as the horizontal axis increases while the vertical remains constant, the color intensity gradually increases, indicating that the subject's state at a historical moment is influenced by the entire historical period, with greater influence from more distant moments, highlighting the subject's anticipation of future instances. In row (b), only a few regions in the heatmap are highlighted in red, demonstrating that our designed perception mask effectively filters out distant neighbors during neighbor cluster processing at each historical moment, focusing only on relevant close neighbors and adhering to social norms, thereby reducing unnecessary computational cost. Row (c) depicts the influence weights in the subject-neighbor groups interaction, which, when combined with the $\mathbf{V}US$ to get the final attention score, serving as the foundation for subsequent decoding in the network.

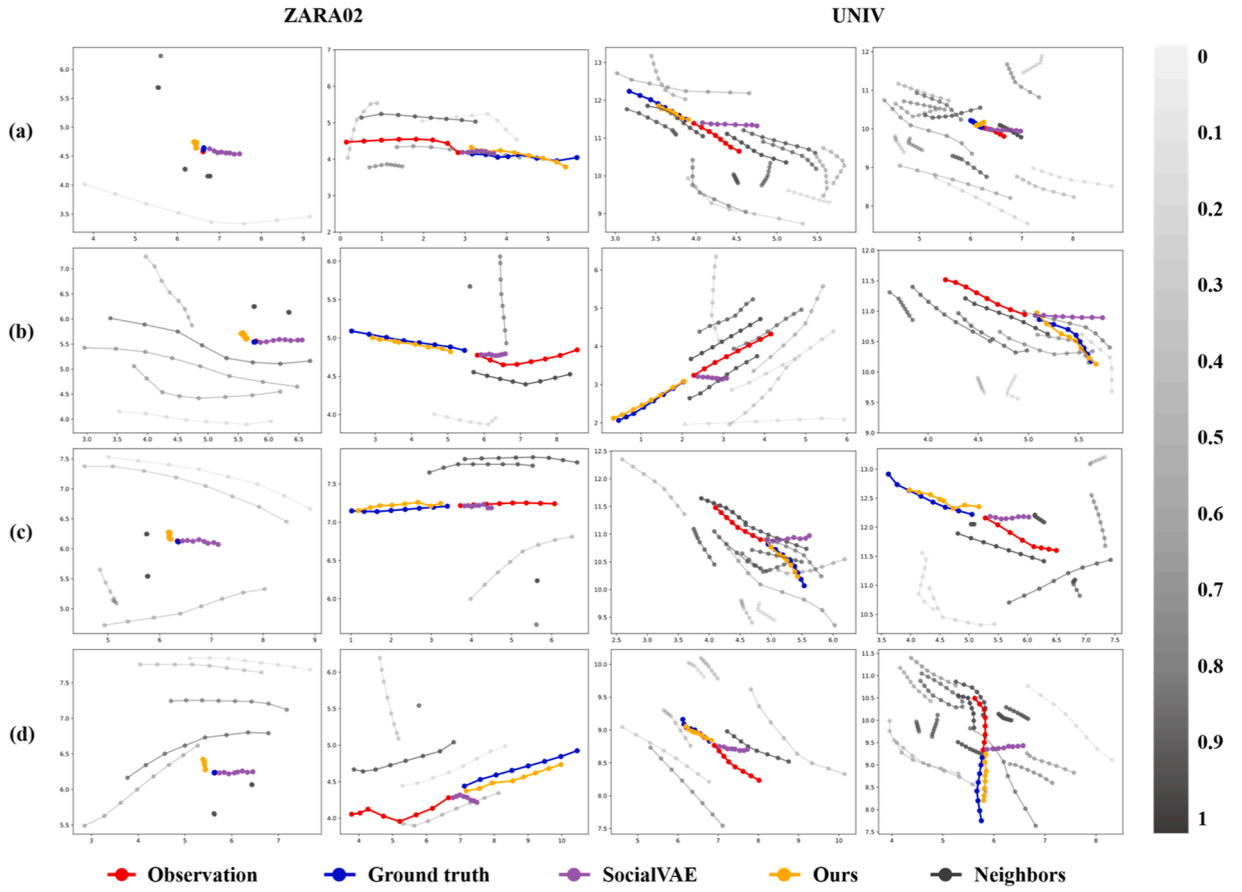

Fig. 9. Trajectory visualization in complex scenarios. We show scenarios with richer social information and a larger number of pedestrians. Neighbors are represented by gray trajectories, and attention scores are characterized by transparency, with lower transparency representing higher attention scores and higher transparency representing lower attention scores.

4.5. Ablation experiment

In order to examine the characteristics of the neighbor perception state tensor, the effects of the two attentions as well as the loss function. We carry out the ablation experiments shown in Table 2 with a network layer number of 2 and keeping the subject state tensor features unchanged. Experiments (1), (2), and (7) investigate the influence of states selection on trajectory prediction results. In (1), only the distance is included, while in (2), the experiments continue by adding the velocity angle θ_{ij}^{t}. Experiment (7) incorporates all statistics mentioned in the methods section. The results show a significant improvement in the prediction outcomes due to states selection, with increases of 50 % and 51 % for the ADE and FDE, respectively. Experiments (3), (4), and (7) primarily examine the impact of attention on trajectory prediction results. Experiments (3), (4), and (7) primarily examine the impact of attention on trajectory prediction results. Experiment (3) includes only the SIE without the NPIE output as input. The SIE makes trajectory predictions based solely on the subject, which contradicts social norms. In experiment (4), NPIE is retained, but the subject's information is ignored, leading to inaccurate predictions. The results indicate greater prediction accuracy when the SIE and NPIE interact. Experiments (5), (6), and (7) focus on the selection of the optimizer. We experiment separately with L_{dis} and L_{angle}, as well as with their joint application. The findings indicate that when employing both loss functions concurrently, the network can effectively learn the distance information between predicted and actual trajectories at each moment, alongside angle information within each time step, resulting in improved experimental outcomes. Additionally, by rectifying angle information, a notable enhancement in FDE performance is observed.

In the first ablation experiment, the best outcomes are achieved within the experimental group (3). Using this group as the baseline, we explored how varying the number of layers in the network affected optimal performance. Table 3 illustrates a noticeable pattern: while increasing the number of network layers, the model's performance does not improve steadily as expected; instead, it demonstrates a degradation trend. Optimal performance is attained when the network depth is set to 2. This pattern suggests that in social scenarios, as network depth increases, the model may excessively focus on learning information between neighbors while overlooking other crucial states. This tendency towards overfitting diminishes the model's ability to generalize, thereby affecting trajectory prediction accuracy. Thus, when designing a network, it is essential to strike a balance between network depth and the model's

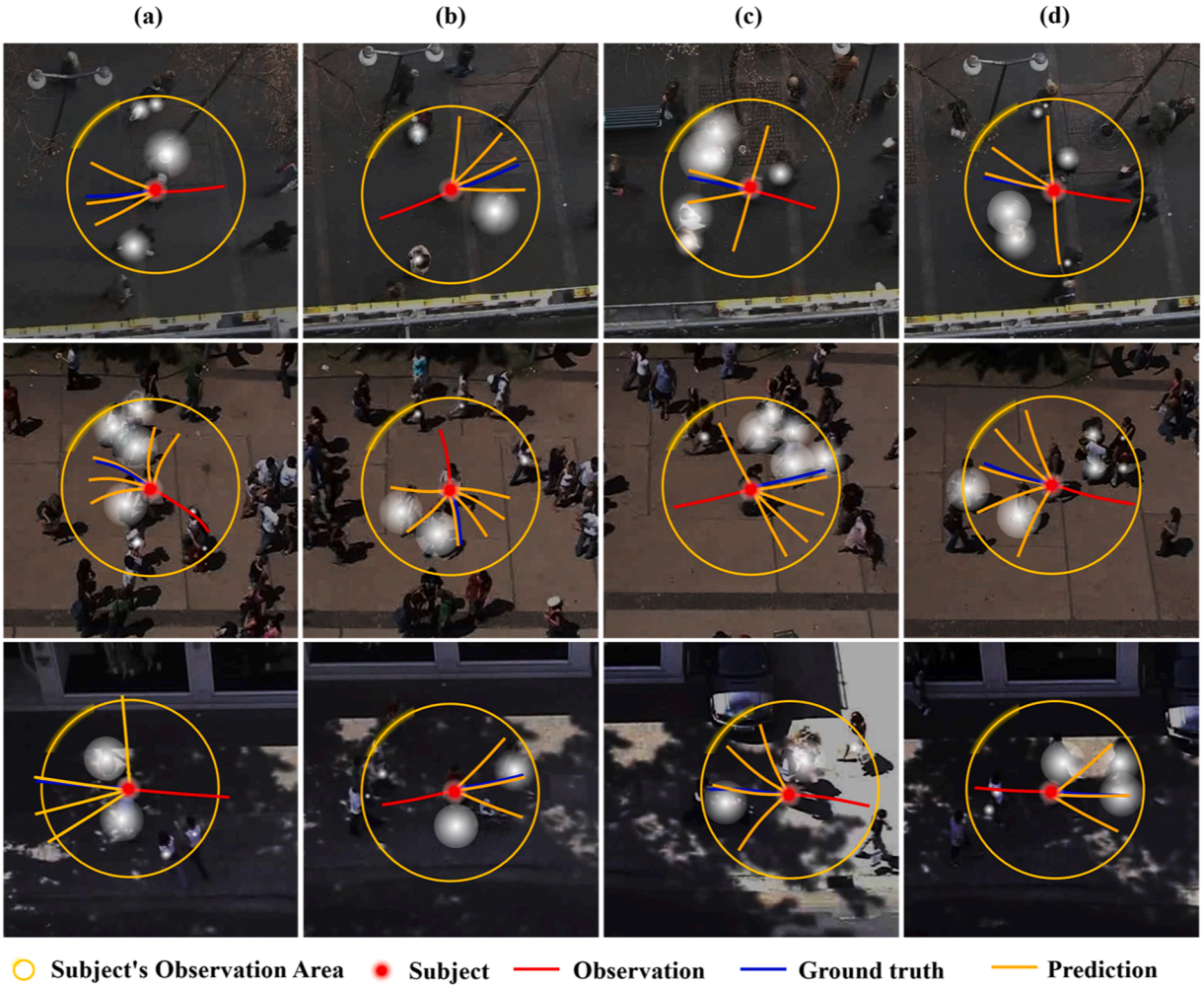

Fig. 10. Attention map within the observation area. The yellow circle is the observation area of the pedestrian, the red dot is the position of the subject itself in the current scenario, the red line segment represents the subject's historical trajectory in the current scenario, the blue line segment represents the real trajectory, and the orange line segment represents the predicted trajectory.

generalization ability to ensure consistent performance across diverse scenarios.

5. Conclusions

We proposed SocialTrans as a novel approach to PTP, which extracted data information from global observations of selected historical periods and could handle an arbitrary number of pedestrians. By modelling the motion states and incorporating future potential information, the state tensor of the subject and the neighbors are then constructed separately. SIE and NPIE are designed for them respectively to realise the internal interaction of social states in social scenarios. To enable SocialTrans to better learn about social information in historical scenarios from a more macroscopic perspective, SIE and NPIE also acted simultaneously throughout the period of observation. Considering the unnecessary computational overhead of dealing with the weak effects caused by distant neighbors, we designed a perception mask in NPIE to perform local processing. The Trajectory Prediction Optimiser makes SocialTrans more accurate in its end-of-period prediction results by fusing distance-angle information, greatly facilitating fast and accurate decision-making for pedestrians. Experimental results on the publicly available datasets ETH/UCY as well as SDD showed that the method outperformed existing methods.

In future work, we need to incorporate more useful information to improve the accuracy of PTP, especially in scenarios where individuals can freely move and change direction at their discretion. This will further demonstrate greater application value in areas such as smart transportation and intelligent manufacturing technology.

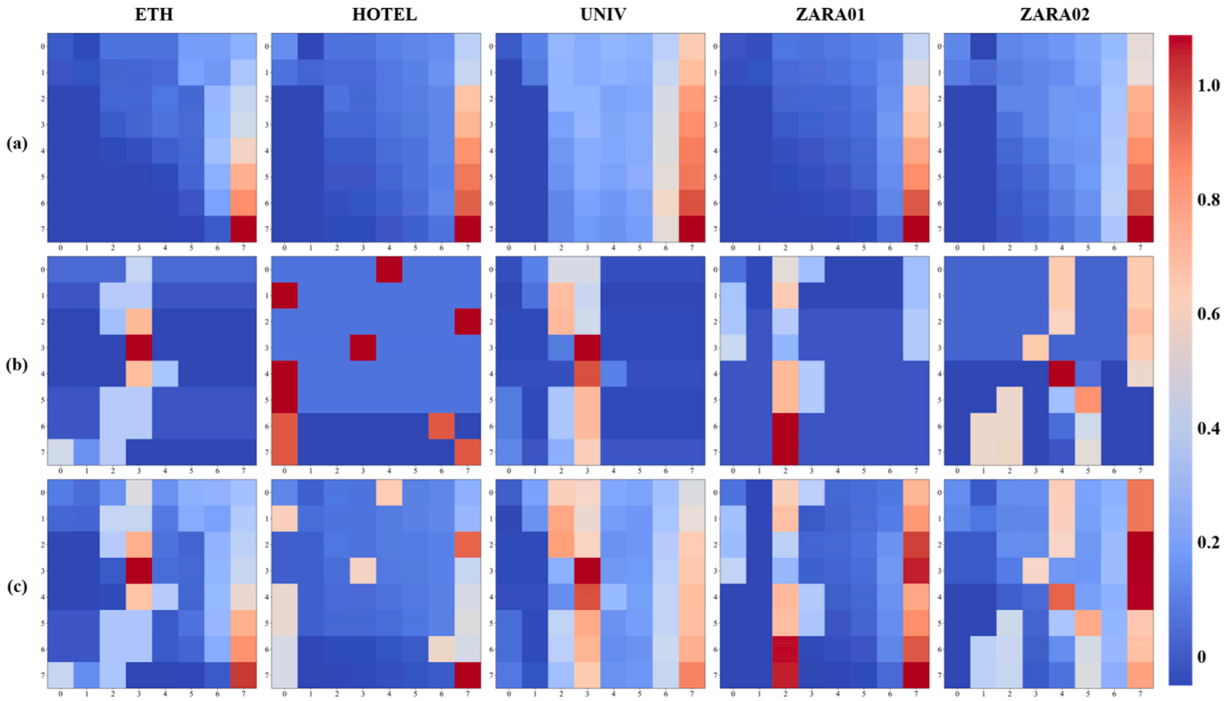

Fig. 11. Social state interaction attention weight heatmap. Both horizontal and vertical coordinates represent historical moments. Row (a) represents the subject state interaction weight heatmap at each historical moment, row (b) represents the neighbor groups state interaction weight heatmap at each historical moment after perceptual masking, and row (c) represents the interaction weight heatmap after (b) is superimposed on (a).

Table 2
Ablation experiment 1: Both ETH/UCY and SDD are carried out on the test dataset. The ETH/UCY results are averaged over the results computed through the five scenarios in meters and the SDD in pixels. The results are reported in the form of ADE/FDE.

	Neighbor States			Attention		Optimizer		Datasets	
	$\left\|\left\|\vec{l}_{ij}^{t}\right\|\right\|$	θ_{ij}^{t}	ob_{ij}^{t}	SIE	NPIE	L_{dis}	L_{angle}	ETH/UCY	SDD
(1)	√	-	-	√	√	√	√	0.29/0.31	5.91/6.95
(2)	√	√	-	√	√	√	√	0.26/0.30	5.12/6.56
(3)	√	√	√	√	-	√	√	0.33/0.38	6.22/7.26
(4)	√	√	√	-	√	√	√	0.31/0.33	5.10/5.98
(5)	√	√	√	√	√	√	-	0.18/0.26	3.95/4.68
(6)	√	√	√	√	√	-	√	0.36/0.42	7.68/8.24
(7)	√	√	√	√	√	√	√	**0.15/0.15**	**3.30/4.11**

Table 3
Ablation experiment 2: The effect of the number of network layers on the experimental results.

	Layers			Datasets	
	2	4	6	ETH/UCY	SDD
(1)	√			**0.15/0.15**	**3.30/4.11**
(2)		√		0.22/0.24	5.61/6.86
(3)			√	0.71/0.88	9.67/10.24

Funding

This work was supported by the National Natural Science Foundation of China (52202417); Fundamental Research Funds for the Central Universities (NS2024030); China Postdoctoral Science Foundation (2022TQ0155, 2022M721605).

CRediT authorship contribution statement

Huang Yujie: Validation, Supervision, Project administration, Investigation. **Zhao Xiaodong:** Writing – original draft, Software, Methodology. **Fang Guoyu:** Writing – original draft, Resources, Methodology. **Chen Kai:** Writing – original draft, Supervision, Methodology, Investigation.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Kai Chen reports financial support was provided by National Natural Science Foundation of China. Kai Chen reports financial support was provided by Fundamental Research Funds for the Central Universities. Kai Chen reports financial support was provided by China Postdoctoral Science Foundation. Kai Chen reports financial support was provided by Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Data will be made available on request.

References

[1] X. Pei, J.B. Hayet, and I. Karamouzas, Socialvae: Human trajectory prediction using timewise latents, in Eur. Conf. Comput. Vis. (ECCV), Cham, Switzerland, Oct., 2022, pp. 511-528.
[2] K. Chen, X. Song, X. Ren, Modeling social interaction and intention for pedestrian trajectory prediction, Phys. Stat. Mech. Its Appl. 570 (2021) 125790.
[3] F. Marchetti, F. Becattini, L. Seidenari, A.D. Bimbo, Mantra: Memory augmented networks for multiple trajectory prediction, Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) (2020) 7143–7150.
[4] C. Xu, W. Mao, W. Zhang, S. Chen, Remember intentions: retrospective-memory-based trajectory prediction, Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) (2022) 6488–6497.
[5] K. Chen, X. Song, D. Han, J. Sun, Y. Cui, X. Ren, Pedestrian behavior prediction model with a convolutional LSTM encoder–decoder, Phys. Stat. Mech. Its Appl. 560 (2020) 125132.
[6] R. Korbmacher, A. Tordeux, Review of pedestrian trajectory prediction methods: comparing deep learning and knowledge-based approaches, IEEE Trans. Intell. Transp. Syst. 23 (12) (2022) 24126–24144.
[7] D. Helbing, I. Farkas, T. Vicsek, Simulating dynamical features of escape panic, Nature 407 (9) (2000) 487–491.
[8] Y. Luo, P. Cai, A. Bera, D. Hsu, W.S. Lee, D. Manocha, PORCA: Modeling and planning for autonomous driving among many pedestrians, IEEE Robot. Autom. Lett. 3 (4) (2018) 3418–3425.
[9] J. Van Den Berg, S.J. Guy, M. Lin, and D. Manocha, Reciprocal n-body collision avoidance, in Robotics Research: The 14th Int. Symp. ISRR, Berlin, Heidelberg, 2011, pp. 3-19.
[10] B.T. Morris, M.M. Trivedi, Trajectory learning for activity understanding: unsupervised, multilevel, and long-term adaptive approach, IEEE Trans. Pattern Anal. Mach. Intell. 33 (11) (2011) 2287–2301.
[11] S. Hoogendoorn, P.H.L. Bovy, Simulation of pedestrian flows by optimal control and differential games, Optim. Control Appl. Methods 24 (3) (2003) 153–172.
[12] W. Yu, D. Helbing, Game theoretical interactions of moving agents. Simulating Complex Systems by Cellular Automata, Berlin, Germany, Springer Berlin Heidelberg, Heidelberg, 2010, pp. 219–239.
[13] S. Bouzat, M.N. Kuperman, Game theory in models of pedestrian room evacuation, Phys. Rev. E 89 (3) (2014) 032806.
[14] S.P. Hoogendoorn, W. Daamen, Y. Shu, H. Ligteringen, Modeling human behavior in vessel maneuver simulation by optimal control and game theory, Transp. Res. Rec. 2326 (1) (2013) 45–53.
[15] X. Zheng, Y. Cheng, Conflict game in evacuation process: a study combining cellular automata model, Phys. A: Stat. Mech. Appl. 390 (6) (2011) 1042–1050.
[16] S. Yi, H. Li, X. Wang, Understanding pedestrian behaviors from stationary crowd groups, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (2015) 3488–3496.
[17] Y. Zhang, L. Qin, H. Yao, Q. Huang, Abnormal crowd behavior detection based on social attribute-aware force model, Proc. 19th IEEE Int. Conf. Inf. Process. (2012) 2689–2692.
[18] K. Yamaguchi, A.C. Berg, L.E. Ortiz, T.L. Berg, Who are you with and where are you going? Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (2011) 1345–1352.
[19] M. Moussaïd, N. Perozo, S. Garnier, D. Helbing, G. Theraulaz, The walking behaviour of pedestrian social groups and its impact on crowd dynamics, PLoS One 5 (4) (2010) e10047.
[20] S. Yi, H. Li, X. Wang, Understanding pedestrian behaviors from stationary crowd groups, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (2015) 3488–3496.
[21] X. Song, K. Chen, Pedestrian trajectory prediction based on deep convolutional LSTM network, IEEE Trans. Intell. Transp. Syst. 22 (6) (2021) 3285–3302.
[22] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, S. Savarese, Social LSTM: human trajectory prediction in crowded spaces, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (2016) 961–971.
[23] H. Manh, and G. Alaghband, Scene-lstm: A model for human trajectory prediction, 2018, arXiv:1808.04018.
[24] Y. Xu, Z. Piao, S. Gao, Encoding crowd interaction with deep neural network for pedestrian trajectory prediction. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Salt Lake City, UT, Jun. 2018, pp. 5275–5284.
[25] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, A. Alahi, Social GAN: socially acceptable trajectories with generative adversarial networks. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 2255–2264.
[26] Y. Ma, X. Zhu, S. Zhang, R. Yang, W. Wang, D. Manocha, TrafficPredict: trajectory prediction for heterogeneous traffic-agents, Proc. AAAI Conf. Artif. Intell. 33 (01) (2019).
[27] P. Zhang, W. Ouyang, P. Zhang, J. Xue, N. Zheng, SR-LSTM: State refinement for LSTM towards pedestrian trajectory prediction. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 12085–12094.
[28] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, S. Savarese, Sophie: an attentive GAN for predicting paths compliant to social and physical constraints. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 1349–1358.
[29] J. Amirian, J.-B. Hayet, J. Pettre, Social ways: learning multi-modal distributions of pedestrian trajectories with GANs. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, Long Beach, CA, USA, Jun. 2019, pp. 2964–2972.
[30] A. Vemula, K. Muelling, J. Oh, Social attention: modeling attention in human crowds, Proc. IEEE Int. Conf. Robot. Autom. (ICRA) (2018) 4601–4607.

[31] C. Yu, X. Ma, J. Ren, H. Zhao, and S. Yi, Spatio-temporal graph transformer networks for pedestrian trajectory prediction, in Proc. Eur. Conf. Comput. Vis. (ECCV), Glasgow, UK, Aug. 23–28, 2020, Part XII, pp. 507-523.
[32] K. Chen, H. Zhu, D. Tang, K. Zheng, Future pedestrian location prediction in first-person videos for autonomous vehicles and social robots, Image Vis. Comput. 134 (2023) 104671.
[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[34] Y. Wang, S. Wang, J. Tang, N.O.'Hare, Y. Chang, B. Li, Hierarchical attention network for action recognition in videos, 2016, arXiv:1607.06416.
[35] K. Chen, X. Song, H. Yuan, X. Ren, Fully convolutional encoder-decoder with an attention mechanism for practical pedestrian trajectory prediction, IEEE Trans. Intell. Transp. Syst. 23 (11) (2022) 20046–20060.
[36] F. Giuliari, I. Hasan, M. Cristani, F. Galasso, Transformer networks for trajectory forecasting, Proc. IEEE Int. Conf. Pattern Recognit. (ICPR) (2021) 10335–10342.
[37] K. Mangalam, H. Girase, S. Agarwal, K.H. Lee, E. Adeli, and J. Malik, It is not the journey but the destination: Endpoint conditioned trajectory prediction, in Proc. Eur. Conf. Comput. Vis. (ECCV), Glasgow, UK, Aug. 2020, Part II, pp. 759-776.
[38] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data, in Proc. Eur. Conf. Comput. Vis. (ECCV), Glasgow, UK, Aug. 2020, Part XVIII, pp. 683-700.
[39] C. Wang, Y. Wang, M. Xu, D.J. Crandall, Stepwise goal-driven networks for trajectory prediction, IEEE Robot. Autom. Lett. 7 (2) (2022) 2716–2723.
[40] Y. Yao, E. Atkins, M. Johnson-Roberson, R. Vasudevan, X. Du, Bitrap: Bi-directional pedestrian trajectory prediction with multi-modal goal estimation, IEEE Robot. Autom. Lett. 6 (2) (2021) 1463–1470.
[41] Y. Yuan, X. Weng, Y. Ou, K.M. Kitani, Agentformer: agent-aware transformers for socio-temporal multi-agent forecasting, Proc. IEEE Int. Conf. Comput. Vis. (ICCV) (2021) 9813–9823.
[42] D.P. Kingma and M. Welling, Auto-Encoding Variational Bayes, 2022, *arXiv*: arXiv:1312.6114.
[43] Z. Cao, H. Gao, K. Mangalam, Q.Z. Cai, M. Vo, J. Malik, Long-term human motion prediction with scene context, Eur. Conf. Comput. Vis. (ECCV (2020) 387–404.
[44] K. Mangalam, Y. An, H. Girase, J. Malik, From goals, waypoints & paths to long term human trajectory forecasting, Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV) (2021) 15233–15242.
[45] L. Ballan, F. Castaldo, A. Alahi, F. Palmieri, S. Savarese, Knowledge transfer for scene-specific motion prediction, Eur. Conf. Comput. Vis. (ECCV) (2016) 697–713.
[46] S. Li, W. Li, C. Cook, C. Zhu, Y. Gao, Independently recurrent neural network (IndRNN): Building a longer and deeper RNN, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (2018) 5457–5466.
[47] J. Li, Y. Wei, X. Liang, J. Dong, T. Xu, J. Feng, S. Yan, Attentive contexts for object detection, IEEE Trans. Multimed. 19 (5) (2016) 944–954.
[48] S. Pellegrini, A. Ess, K. Schindler, L.V. Gool, You'll never walk alone: Modeling social behavior for multi-target tracking, Proc. IEEE 12th Int. Conf. Comput. Vis. (2009) 261–268.
[49] A. Lerner, Y. Chrysanthou, D. Lischinski, Crowds by example, Comput. Graph. Forum 26 (3) (2007) 655–664.
[50] J. Gehring, M. Auli, D. Grangier, D. Yarats, Y.N. Dauphin, Convolutional sequence to sequence learning, Proc. Int. Conf. Mach. Learn. (2017) 1243–1252.
[51] H. Li, A. Rakhlin, A. Jadbabaie, Convergence of Adam under relaxed assumptions, Adv. Neural Inf. Process. Syst. 30 (2024).
[52] A. Robicquet, A. Sadeghian, A. Alahi, S. Savarese, Learning social etiquette: Human trajectory prediction in crowded scenes, Proc. Eur. Conf. Comput. Vis. (ECCV) 2 (4) (2016) 5.
[53] S. Becker, R. Hug, W. Hübner, and M. Arens, An evaluation of trajectory prediction approaches and notes on the Trajnet benchmark, 2018, arXiv:1805.07663.
[54] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, Show, attend and tell: neural image caption generation with visual attention, Proc. Int. Conf. Mach. Learn. (2015) 2048–2057.
[55] R. Wu, X. Zheng, Y. Xu, W. Wu, G. Li, Q. Xu, Modified driving safety field based on trajectory prediction model for pedestrian–vehicle collision, Sustainability 11 (22) (2019) 6254.
[56] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, Proc. Int. Conf. Mach. Learn. (2019) 1597–1607.
[57] X. Chen, H. Fan, R. Girshick, and K. He, Improved baselines with momentum contrastive learning, Mar. 2020, arXiv:2003.04297.
[58] Y. Chang, J. Qi, Y. Liang, E. Tanin, Contrastive trajectory similarity learning with dual-feature attention, Proc. IEEE 39th Int. Conf. Data Eng. (ICDE (2023) 2933–2945.
[59] K. Chen, X. Song, X. Ren, Pedestrian trajectory prediction in heterogeneous traffic using pose keypoints-based convolutional encoder-decoder network, IEEE Trans. Circuits Syst. Video Technol. 31 (5) (2020) 1764–1775.
[60] X. Song, K. Chen, X. Li, et al., Pedestrian trajectory prediction based on deep convolutional LSTM network, IEEE Trans. Intell. Transp. Syst. 22 (6) (2020) 3285–3302.