Engineering 34 (2024) 70–82

Contents lists available at ScienceDirect

Engineering

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

Research
Artificial Intelligence—Article

The Group Interaction Field for Learning and Explaining Pedestrian Anticipation

Xueyang Wang [a,b,c,#], Xuecheng Chen [c,#], Puhua Jiang [c,#], Haozhe Lin [b,d,#], Xiaoyun Yuan [a,b], Mengqi Ji [e], Yuchen Guo [b], Ruqi Huang [c], Lu Fang [a,b,f,*]

[a] Sigma Laboratory, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[b] Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China
[c] Tsinghua Shenzhen International Graduate School, Shenzhen 518055, China
[d] Department of Automation, Tsinghua University, Beijing 100084, China
[e] Institute of Artificial Intelligence, Beihang University, Beijing 100191, China
[f] Zhejiang Future Technology Institute, Yangtze Delta Region Institute of Tsinghua University, Zhejiang, Jiaxing 314033, China

ARTICLE INFO

Article history:
Received 2 July 2022
Revised 10 December 2022
Accepted 9 May 2023
Available online 23 August 2023

Keywords:
Human behavior modeling and prediction
Implicit representation of pedestrian anticipation
Group interaction
Graph neural network

ABSTRACT

Anticipating others' actions is innate and essential in order for humans to navigate and interact well with others in dense crowds. This ability is urgently required for unmanned systems such as service robots and self-driving cars. However, existing solutions struggle to predict pedestrian anticipation accurately, because the influence of group-related social behaviors has not been well considered. While group relationships and group interactions are ubiquitous and significantly influence pedestrian anticipation, their influence is diverse and subtle, making it difficult to explicitly quantify. Here, we propose the group interaction field (GIF), a novel group-aware representation that quantifies pedestrian anticipation into a probability field of pedestrians' future locations and attention orientations. An end-to-end neural network, GIFNet, is tailored to estimate the GIF from explicit multidimensional observations. GIFNet quantifies the influence of group behaviors by formulating a group interaction graph with propagation and graph attention that is adaptive to the group size and dynamic interaction states. The experimental results show that the GIF effectively represents the change in pedestrians' anticipation under the prominent impact of group behaviors and accurately predicts pedestrians' future states. Moreover, the GIF contributes to explaining various predictions of pedestrians' behavior in different social states. The proposed GIF will eventually be able to allow unmanned systems to work in a human-like manner and comply with social norms, thereby promoting harmonious human–machine relationships.

1. Introduction

Understanding pedestrian dynamics is critical in a variety of real-world tasks, such as autonomous driving [1,2], robot navigation [3,4], pedestrian flow analysis [5,6], and crowd evacuation [7,8]. Interestingly, humans have an instinctive ability to anticipate the future actions of other people while navigating in crowded spaces and interacting with other pedestrians [9–13], which permits them to avoid head-on collisions and keep pace with peer partners while maintaining a comfortable distance. As shown in Fig. 1(a), such an ability would allow unmanned systems to work in urban environments intelligently by comprehending and anticipating the actions of pedestrians.

In the past decades, pedestrian anticipation has been modeled using bidirectional flow [13,14], cellular automation [10,15], and time to collision [12,16,17] to simulate collective behaviors. Recently, machine learning technology has been utilized for this purpose, allowing the future states of pedestrians to be forecasted [17–23]. In essence, the above methods model each individual's behavior in collision avoidance without considering group-related social behaviors. However, humans are naturally social beings who gather to interact socially and thus form social groups [24,25]; for example, up to 70% of the observed pedestrians on a street are in groups [26]. Pedestrians conform to expected social

* Corresponding author.
E-mail address: fanglu@tsinghua.edu.cn (L. Fang).
# These authors contributed equally to this work.

https://doi.org/10.1016/j.eng.2023.05.020

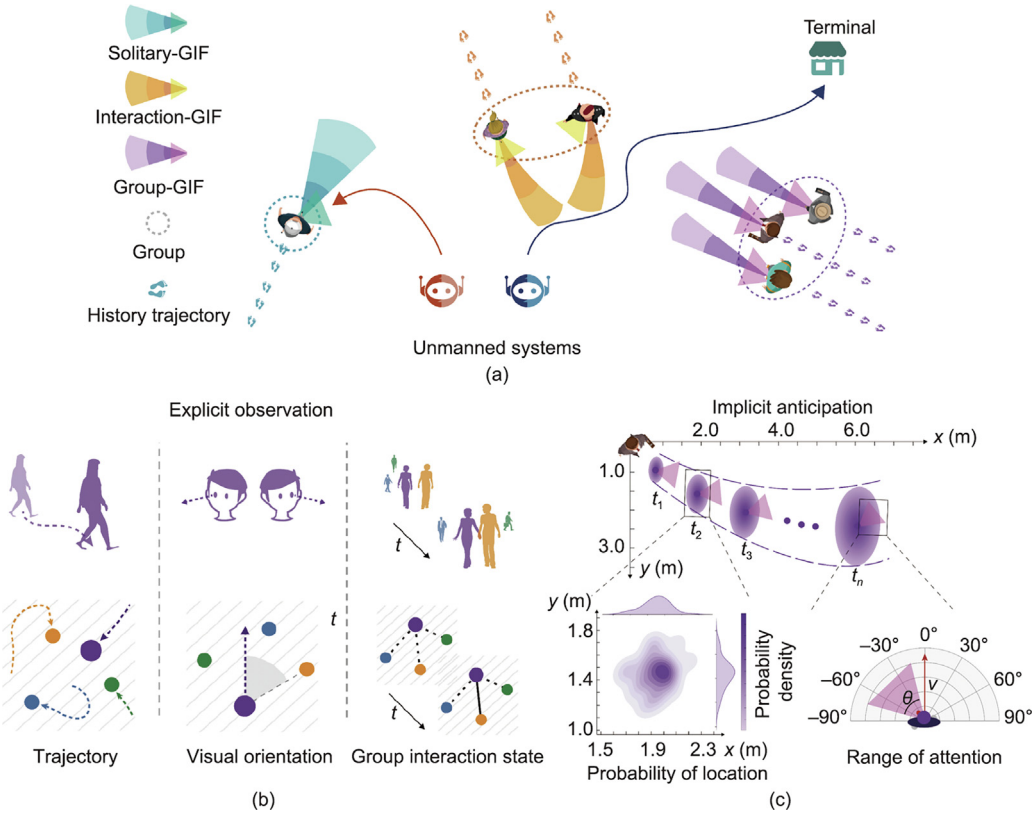

Fig. 1. The group interaction field (GIF). (a) The GIF represents implicit pedestrian anticipation; it consists of a proxemics field and an attention field, estimated from explicit observations. (b) Explicit observations consist of trajectory, visual orientation, and observable group interaction state. (c) The proxemics field and the attention field are represented by a sequence of two-dimensional (2D) probabilistic distribution maps and a sequence of angular ranges, respectively. Two representative applications are demonstrated in part (a); the GIF can help unmanned systems to either avoid disturbing pedestrians or attract pedestrians' attention. t_1, t_2,..., t_n: timestamp from the predication sequence; n: nth timestamp; v: walking velocity; θ: angular range of the attention field.

norms in groups and act accordingly under the influence of group neighbors [27], where intra-/inter-group interactions are considered to be critical influencers of pedestrians' social cognition [27,28] and behavior patterns [29,30]. To model group or interaction information, state-of-the-art graph neural network (GNN) methods have been utilized for an understanding of pedestrian/agent dynamics [9,27,31–35]. However, for pedestrians, the influence of group behaviors is not only diverse but also subtle, and different group relationships or interaction states will have very different impacts on pedestrians' future states. For example, a family group (e.g., mother and daughter) and a tour group usually show quite different behaviors under similar scenarios, as the attention of children is less focused than that of adults. These subtle differences cannot be well modeled by simple relationship or interaction graphs. Because they fail to distinguish between different group relationships among pedestrians, existing methods are insufficient for accurately predicting the differences in pedestrian anticipation influenced by group behaviors [13,19,23,27,31,34–37].

In complex scenes, it is important yet challenging to understand the influence of group relationships and social interactions on pedestrian behaviors. For such contexts, we propose the group interaction field (GIF), a novel group-aware representation, to quantify implicit pedestrian anticipation. More specifically, the GIF consists of a proxemics field and an attention field, which respectively represent pedestrian anticipation using the probability fields of pedestrians' future locations and their attention orientations. Moreover, we tailor GIFNet to estimate the GIF from explicit multidimensional observations, including the trajectory, visual orientation, and group interaction state. GIFNet can quantify the diverse and subtle influence of group behaviors by formulating a group interaction graph with propagation and group attention that is adaptive to the group size and dynamic interaction states. Our main contributions are threefold:

- We propose the GIF, a group-aware representation of pedestrian anticipation. It consists of a proxemics field and an attention field, which represent the variation of pedestrian anticipation, and thus delivers a comprehensive understanding of the social nature of pedestrians.
- We tailor GIFNet to estimate the GIF; taking explicit observations into consideration, GIFNet uses the advantages of long short-term memory (LSTM) and graph attention network (GAT) to learn implicit spatiotemporal representation and estimate the GIF.
- Extensive validation in various real-world scenarios shows that the GIF can effectively represent changes in pedestrian anticipation under the prominent impact of group behaviors and accurately predict pedestrians' future states.

2. The GIF

As an estimation of implicit pedestrian anticipation, the GIF consists of a proxemics field and an attention field, which respectively represent predictions of pedestrians' future location and visual attention. The GIF is estimated by means of GIFNet from explicit observations of pedestrians, including their trajectory, visual orientation, and state of group interaction (Fig. 1(b)). We generated pedestrian data from the PANDA dataset [38], which consists of large-scale natural outdoor scenes with a diversity of scenarios, as well as pedestrian density, trajectory distribution,

and group activities. The proxemics field is a sequence of two-dimensional (2D) probabilistic distribution maps denoting the future locations of the pedestrian of interest (Fig. 1(c)) with a timespan T and temporal resolution R. Similarly, the attention field is a sequence of angular ranges θ, representing the pedestrian's possible orientation and range of visual attention. More formally, given a timestamp from the observation sequence $t \in \{1, \ldots, T_{\text{end}}\}$, with the ending timestamp T_{end}, of the pedestrian of interest i, the GIF is defined as $\text{GIF}_i^T = \left[P_i^T, A_i^T\right]$, with the proxemics field P_i^T and attention field A_i^T.

As shown in Fig. 1(a), the GIFs of solitary pedestrians (cyan), grouped pedestrians without interaction (purple), and grouped pedestrians with interaction (orange) have apparent differences: The single pedestrian has a long and wide proxemics field, while the grouped pedestrians without interaction have shorter and narrower fields, and the grouped pedestrians with interaction tend to approach each other closely. As it can predict pedestrians' future location and attention orientation, the GIF has great potential in unmanned system applications. Fig. 1(a) shows two representative applications of the GIF. The proxemics field can help an unmanned system (blue) plan its path to avoid disturbing pedestrians, while the attention field can guide an unmanned system (red) to approach a pedestrian from the orientation of attention.

3. GIFNet

To accurately estimate the GIF, we tailor GIFNet, as illustrated in Fig. 2(a). GIFNet takes three explicit observations as inputs—namely, the trajectory of the pedestrian of interest T_{p}, the visual orientation of the pedestrian of interest F_{p}, and the neighbor trajectories T_{n} in an interaction graph I_{t}, with timestamp t, and outputs the GIF of the pedestrian of interest. Given the pedestrian of interest (purple), the remaining pedestrians in the same group (other colors) are denoted as that pedestrian's neighbors. More specifically, the group interaction graph I_{t} is a graph sequence for organizing the group interaction state, whose edges represent whether the pedestrian of interest is interacting with neighbors at each timestep.

GIFNet consists of four modules: ① a trajectory encoder that models the historical trajectory of the pedestrian of interest, ② an visual orientation encoder that models the pedestrian of interest's visual orientation information, ③ the GIF-GAT, which models the interaction information between the pedestrian of interest and that pedestrian's neighbors, and ④ a visual orientation decoder and proxemics decoder that respectively generate an estimation of the proxemics field and of the attention field of the pedestrian of interest. In GIFNet, three encoders composed of a fully connected (FC) layer and an LSTM unit are used to extract features from T_{p}, F_{p}, and T_{n}. For the neighbor trajectories T_{n}, the encoder produces two embedding vectors ($\boldsymbol{m}_{\text{a}}^j$ and $\boldsymbol{m}_{\text{r}}^j$) for the jth neighbor, encoding the features of the neighbor's absolute displacement and the displacement relative to the pedestrian of interest, respectively (Fig. 2(b)). The group interaction graph I_{t} and the features of the neighbor trajectories T_{n} are further processed by means of a graph attention module (the GIF-GAT; Fig. 2(c)). For each timestep t, an FC layer is used to calculate the weights of the neighbors from the relative displacement feature of the neighbors. The weights are multiplied by the group interaction graph I_{t} to obtain the final weight α_j for the jth neighbor. The absolute displacement features of the neighbors ($\boldsymbol{m}_{\text{a}}^j$) are then summed with weighting, using α_j, as the final neighbor embedding vector. In this way, GIFNet propagates the influence of the neighbors and the group interactions through the graph to learn an embedding feature vector. Finally, the embedding feature vectors of the four kinds of explicit observations are input to the decoders for estimating the proxemics field and the attention field (Fig. 2(d)). For the proxemics decoder, a Gaussian sampling module is added to learn the uncertainty of the proxemics field and produce a sequence of probability distribution maps representing the pedestrian of interest's future location. In the following, we will elaborate the design of the trajectory encoder, the visual orientation encoder, GIF-GAT, and the decoders.

3.1. Trajectory encoder

The purpose of the trajectory encoder is to encode the historical trajectory information and generate a trajectory embedding. The trajectory encoder consists of LSTM_{X}. The past trajectory information X_i of pedestrian of interest i is represented by the ordered set of the pedestrian's relative displacement to the previous timestep (Fig. 2(b)) and is formed as follows:

$$X_i = \left\{\Delta x_i^1, \ldots, \Delta x_i^{T_{\text{end}}}\right\} \tag{1}$$

$$\Delta x_i^t = x_i^t - x_i^{t-1} \tag{2}$$

where x_i^t is the spatial location of person of interest at timestamp t.

For the timesteps $t = \{1, \ldots, T_{\text{end}}\}$, we perform the following update operation to embed the relative displacement into a fixed-length vector $\boldsymbol{e}_i^t$ corresponding to the FC layer in Fig. 2(a):

$$\boldsymbol{e}_i^t = \phi\left(\Delta x_i^t; \boldsymbol{W}_{\text{ee}}\right) \tag{3}$$

Then, the embedding vector is used as input to the LSTM cell, as follows:

$$\boldsymbol{m}_{\text{X}}^t(i) = \text{LSTM}_{\text{X}}\left(\boldsymbol{m}_x^{t-1}(i), \boldsymbol{e}_i^t; \boldsymbol{W}_{\text{X}}\right) \tag{4}$$

where the function ϕ is the FC layer to embed the past trajectory information of pedestrian i, $\boldsymbol{W}_{\text{ee}}$ is the embedding weight, $\boldsymbol{m}_{\text{X}}^t(i)$ is the hidden state of the LSTM_{X} at timestep t, and $\boldsymbol{W}_{\text{X}}$ is the weight of the LSTM_{X} cell. These parameters are shared among all the pedestrians in the scene.

3.2. Visual orientation encoder

The purpose of the visual orientation encoder is to encode the historical visual orientation information and generate a visual orientation embedding. The past visual orientation information $\boldsymbol{A}_i$ of pedestrian of interest i is represented by the ordered set of the pedestrian's orientation a_i^t in a unit vector and is formed as follows:

$$\boldsymbol{A}_i = \left\{a_i^1, \ldots, a_i^{T_{\text{end}}}\right\} \tag{5}$$

$$a_i^t = \left(\cos\theta_1^t, \sin\theta_1^t\right) \tag{6}$$

where θ_1 is the inner angle of visual orientation concerning the forward orientation. Similar to the trajectory encoder module, the visual attention sequence $\boldsymbol{A}_i$ with the hidden state $\boldsymbol{m}_{\text{o}}^t(i)$ is fed into the visual orientation encoder LSTM_{o}. The operation is as follows:

$$\boldsymbol{e}_i^t = \phi\left(\Delta a_i^t; \boldsymbol{W}_{\text{ee}}\right) \tag{7}$$

$$\boldsymbol{m}_{\text{o}}^t(i) = \text{LSTM}_{\text{o}}\left(\boldsymbol{m}_{\text{o}}^{t-1}(i), \boldsymbol{e}_i^t; \boldsymbol{W}_{\text{o}}\right) \tag{8}$$

where $\boldsymbol{W}_{\text{o}}$ is the weight of the LSTM_{o} cell. For simplicity, we reuse the notations of ϕ and $\boldsymbol{W}_{\text{ee}}$ to represent the embedding function and the embedding weight and hidden state, respectively. The final vector $\boldsymbol{m}_{\text{o}}^{T_{\text{end}}}(i)$ is the ensemble of the information from the visual orientation of pedestrian of interest i.

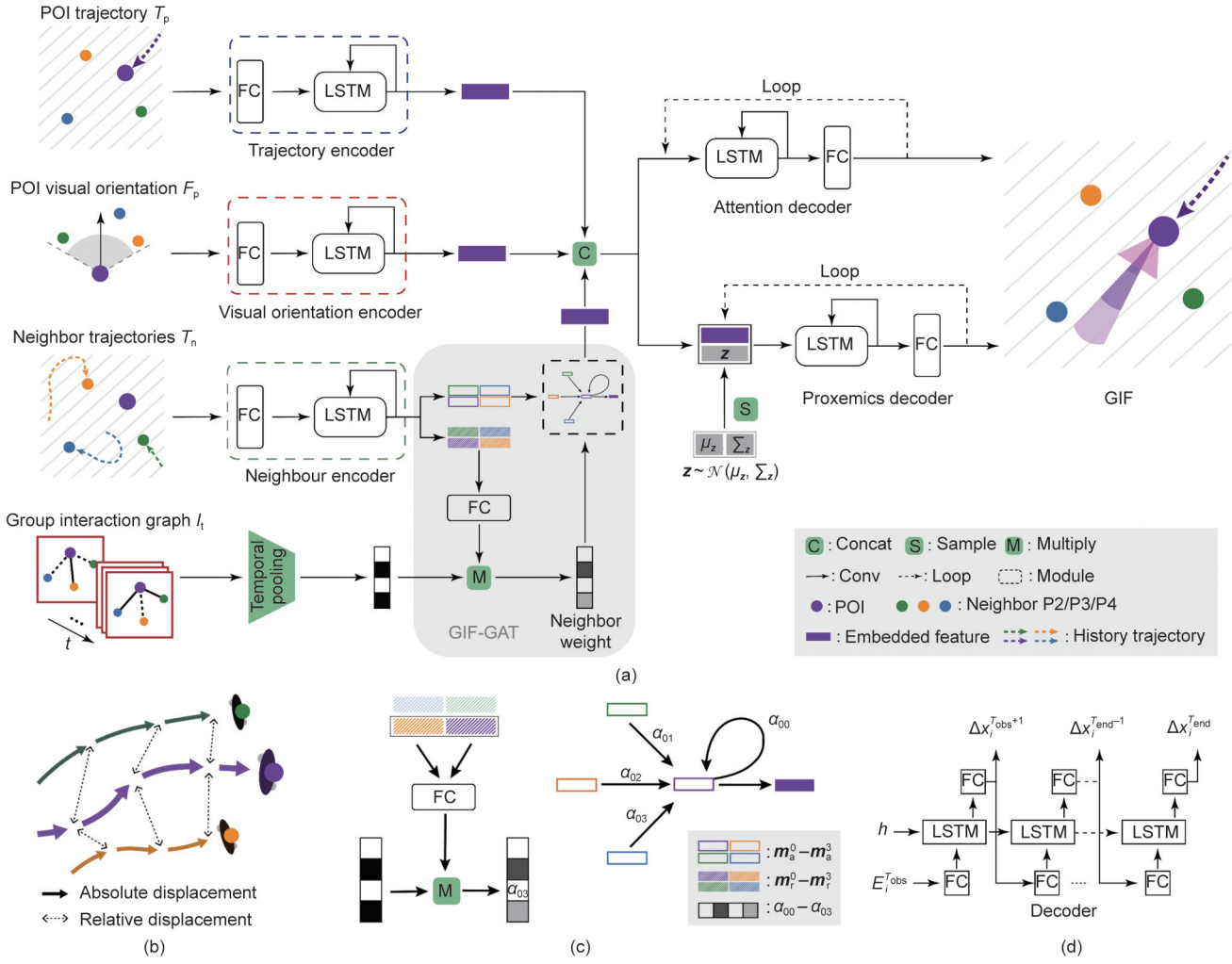

Fig. 2. GIFNet. (a) Network structure of GIFNet, including three encoders (trajectory, visual orientation, and neighbor), GIF-GAT (components in the grey area), and two decoders (proxemics and attention). (b) Illustration of the relative displacement (i.e., the relative neighbor location to the pedestrian of interest) and the absolute displacement (i.e., the neighbors' self-displacement). (c) Network structure of GIF-GAT, where GIF-GAT takes the neighbors' relative embeddings (filled rectangles, $\boldsymbol{m}_r^j$), the neighbors' absolute embeddings (hollow rectangles, $\boldsymbol{m}_a^j$), and the dynamic group interaction graph as input and outputs a fixed-length embedding representing the influence of all group neighbors on the pedestrian of interest. (d) Loop structure in the proxemics decoder and attention decoder for predicting the sequence of future states. POI: pedestrian of interest; FC: fully connected; T_p: observed POI trajectory; F_p: observed POI visual orientation; T_n: observed neighbor trajectories; I_t: observed group interaction graph; $\boldsymbol{z}$: Gaussian noise vector; $\mathcal{N}$: normal distribution; $\boldsymbol{\mu_z}$: mean value of $\boldsymbol{z}$; $\Sigma_{\boldsymbol{z}}$: variance of $\boldsymbol{z}$; conv: convolution; α_{0j}: weight for the jth neighbor; $\boldsymbol{m}_a^j$ and $\boldsymbol{m}_r^j$: two embedding vectors for the jth neighbor; h: hidden states of LSTM; T_{obs}: any one observed timestamp; $E_i^{T_{obs}}$: embeddings of the observed timestamp of person of interest; Δx_i^t: pedestrian's predicted relative displacement to the previous timestep.

3.3. GIF-GAT

For efficiency and simplicity, we adopt a mechanism similar to the trajectory encoder to encode the neighbor trajectories. For the pedestrian of interest i, as shown in Fig. 2(b), in addition to the displacement of each neighbor j to the previous timestep as $\Delta x_j^t = x_j^t - x_j^{t-1}$, we calculate the relative location of each neighbor j in relation to the pedestrian of interest i at each timestep; that is, $\Delta x_{ij}^t = x_i^t - x_j^t$. We encode the neighbor location in both Δx_j^t and Δx_{ij}^t, which represent the absolute displacement and relative displacement, respectively. The operations are as follows:

$$\boldsymbol{e}_i^t = \phi\left(\Delta x_j^t; \boldsymbol{W}_{ee}\right) \tag{9}$$

$$\boldsymbol{e}_{ij}^t = \phi\left(\Delta x_{ij}^t; \boldsymbol{W}_{ee}\right) \tag{10}$$

Then, by feeding the corresponding vectors to the neighbor encoder, we obtain two distance-sensitive context embeddings: the neighbor's relative embedding $\boldsymbol{m}_r^t(j)$ and the neighbor's absolute embedding $\boldsymbol{m}_a^t(j)$. The operations are as follows:

$$\boldsymbol{m}_a^t(i) = \mathrm{LSTM}_E\left(\boldsymbol{m}_a^{t-1}(i), \boldsymbol{e}_i^t; \boldsymbol{W}_a\right) \tag{11}$$

$$\boldsymbol{m}_r^t(i) = \mathrm{LSTM}_E\left(\boldsymbol{m}_r^{t-1}(i), \boldsymbol{e}_{ij}^t; \boldsymbol{W}_a\right) \tag{12}$$

where $\boldsymbol{W}_a$ represents the weight of the correspondingly LSTM cell.

We use a GAT as a sharing mechanism to aggregate the information on interactions between the pedestrian of interest and that pedestrian's neighbors. As shown in Fig. 2, we consider the pedestrians in a scene as nodes and use edges on the graph to represent information on human–human interaction. The GAT is constructed by stacking graph attention layers. The group interaction graph of

the pedestrian of interest i is represented by a sequence of dummy variables, as follows:

$$I_{\mathrm{t}}^{i}=\left\{\beta_{ij}^{t}|j\in\{1,\ldots,D_i\},t\in\{1,\ldots,T_{\mathrm{end}}\}\right\} \tag{13}$$

where β_{ij}^{t} is the dummy variable indicating the existence of interaction between the pedestrian of interest i and group neighbor j, and D_i is the number of group neighbors of the pedestrian of interest i. We adopt temporal pooling for β_{ij}^{t} to generate a pooled context vector $\boldsymbol{C}_{ij}$, which is composed of the interaction information across the observation period; that is, $\boldsymbol{C}_{ij}=\frac{1}{T}\sum_{i}^{T}\beta_{ij}^{t}$.

Let $\boldsymbol{m}_{\mathrm{r}}^{T_{\mathrm{end}}}(j)$ denote the final relative embedding and $\boldsymbol{m}_{\mathrm{a}}^{T_{\mathrm{end}}}(j)$ denote the final absolute embedding of neighbor j. In the observation period, $\boldsymbol{m}_{\mathrm{a}}^{T_{\mathrm{end}}}(j)$ is fed to the graph attention layer. The coefficients in the attention mechanism of the node pair (i,j) can be computed by multiplying $\boldsymbol{m}_{\mathrm{r}}^{T_{\mathrm{end}}}(j)$ and $\boldsymbol{C}_{ij}$ as follows:

$$\alpha_{ij}=\boldsymbol{m}_{\mathrm{r}}^{T_{\mathrm{end}}}(i)\cdot\boldsymbol{C}_{ij} \tag{14}$$

The output of one graph attention layer for node i (pedestrian of interest i) is given by the following:

$$\widehat{\boldsymbol{m}}_{\mathrm{a}}^{T_{\mathrm{end}}}(i)=\sigma\left(\sum\nolimits_{j\in N_i}\alpha_{ij}\boldsymbol{W}\boldsymbol{m}_{\mathrm{a}}^{T_{\mathrm{end}}}(j)\right) \tag{15}$$

where σ is a nonlinear function and N_i represents the neighbors of node i. $\boldsymbol{W}\in R^{F'\times F}$ is the parameter matrix of a shared linear projection that is applied to each neighbor separately (F is the dimension of the input, and F' is the dimension of the output). In addition, $\widehat{\boldsymbol{m}}_{\mathrm{a}}^{T_{\mathrm{end}}}$ is a fixed-length embedding for the pedestrian of interest i for the observed time, representing the influence of all neighbors on the pedestrian of interest.

3.4. Proxemics and attention decoder

We use the decoders to generate the proxemics field and attention field conditioned on $\boldsymbol{E}_t(i)=[\boldsymbol{E}_{\mathrm{p}},\boldsymbol{E}_{\mathrm{v}},\boldsymbol{E}_{\mathrm{n}}]$, where $\boldsymbol{E}_{\mathrm{p}}$, $\boldsymbol{E}_{\mathrm{v}}$, and $\boldsymbol{E}_{\mathrm{n}}$ are the embeddings of the trajectory, visual orientation, and neighbors' influences, respectively:

$$\boldsymbol{E}_{\mathrm{p}}=\phi\left(\boldsymbol{m}_{x}^{T_{\mathrm{end}}}(i);\boldsymbol{W}_{\mathrm{ee}}\right) \tag{16}$$

$$\boldsymbol{E}_{\mathrm{v}}=\phi\left(\boldsymbol{m}_{\mathrm{o}}^{T_{\mathrm{end}}}(i);\boldsymbol{W}_{\mathrm{ee}}\right) \tag{17}$$

$$\boldsymbol{E}_{\mathrm{n}}=\phi\left(\widehat{\boldsymbol{m}}_{\mathrm{a}}^{T_{\mathrm{end}}}(i);\boldsymbol{W}_{\mathrm{ee}}\right) \tag{18}$$

Then, we directly concatenate a noise vector $\boldsymbol{z}$ sampled from a Gaussian distribution and the context embeddings $\boldsymbol{E}_t(i)$ as the input for the proxemics decoder $\mathrm{LSTM}_{\mathrm{dec}}^{\mathrm{p}}$:

$$h_{\mathrm{p}}^{T_{\mathrm{end}}+1}(i)=\mathrm{LSTM}_{\mathrm{dec}}^{\mathrm{p}}\left(h_{\mathrm{p}}^{T_{\mathrm{end}}}(i),[\boldsymbol{E}_t(i),\boldsymbol{z}];\boldsymbol{W}_{\mathrm{dec}}^{\mathrm{p}}\right) \tag{19}$$

$$\Delta x_i^{T_{\mathrm{end}}+1}=\phi\left(h_{\mathrm{p}}^{T_{\mathrm{end}}+1}(i);\boldsymbol{W}_{\mathrm{ee}}\right) \tag{20}$$

Moreover, the attention field of the pedestrian of interest i is updated using the attention decoder $\mathrm{LSTM}_{\mathrm{dec}}^{\mathrm{a}}$:

$$h_{\mathrm{a}}^{T_{\mathrm{end}}+1}(i)=\mathrm{LSTM}_{\mathrm{dec}}^{\mathrm{a}}\left(h_{\mathrm{a}}^{T_{\mathrm{end}}}(i),[\boldsymbol{E}_t(i),\boldsymbol{z}];\boldsymbol{W}_{\mathrm{dec}}^{\mathrm{a}}\right) \tag{21}$$

$$a_i^{T_{\mathrm{end}}+1}=\phi\left(h_{\mathrm{a}}^{T_{\mathrm{end}}+1}(i);\boldsymbol{W}_{\mathrm{ee}}\right) \tag{22}$$

where $\Delta x_i^{T_{\mathrm{end}}+1}$ and $a_i^{T_{\mathrm{end}}+1}$ are respectively the location and the visual orientation of the pedestrian of interest i at $T_{\mathrm{end}}+1$. We use the notations $h_{\mathrm{a}}^{t}(i)$ and $h_{\mathrm{p}}^{t}(i)$ to represent the hidden state of the proxemics decoder and of the attention decoder, respectively, and use $\boldsymbol{W}_{\mathrm{dec}}^{\mathrm{p}}$ and $\boldsymbol{W}_{\mathrm{dec}}^{\mathrm{a}}$ to represent the embedding weight of the proxemics decoder and of the attention decoder, respectively.

4. Experiments

4.1. Experimental settings

4.1.1. Dataset

The performance of our models was evaluated on the PANDA dataset [38]. The videos in the PANDA dataset are captured by gigapixel cameras, and each video frame contains hundreds to thousands of pedestrians, with rich group interaction information. As our method only requires the trajectories, visual orientations, and group interaction information, we extracted this information from the PANDA labels and formed a new dataset with 21 704 trajectories. We divided the trajectories into training, testing, and validation sets, with 15 511, 3052, and 3141 trajectories, respectively. Next, we computed a homography matrix to map images to the top view in order to obtain the locations of the pedestrians in world coordinates.

Unlike the existing group-based trajectory-prediction datasets [22], each group was assigned several category labels, denoting the kinds of group relationships (i.e., acquaintance and family) and interaction states (i.e., no interaction, non-physical interaction, and physical interaction). For example, eye contact, body language such as hand waving, and talking are non-physical interactions, while holding hands is a type of physical interaction. Group relationship information is identified through the interactions and characteristics of the members, such as appearance, gender, age, and exchanges.

4.1.2. Evaluation metrics

During the test time, we made k predictions of the future position of the pedestrian of interest i; we set $k=20$. Then, we applied a Gaussian model to fit the predicted locations for all k predictions and then sampled the point with the highest probability as the optimal predicted location $\widehat{Y}_i^{\mathrm{p}}(t)$, which was calculated as follows:

$$\widehat{Y}_i^{\mathrm{p}}(t)=\max\nolimits_{\Delta x_i^t}P(\Delta x_i^t) \tag{23}$$

where $P(\Delta x_i^t)$ is the fitted Gaussian model for all predicted locations of pedestrian i at time t. We used the average displacement error (ADE) [19] and the final displacement error (FDE) [38] to evaluate the predicted trajectory as follows:

$$\mathrm{ADE}=\frac{1}{N}\sum\nolimits_{t=1}^{N}\left|Y_i^{\mathrm{p}}(t)-\widehat{Y}_i^{\mathrm{p}}(t)\right| \tag{24}$$

$$\mathrm{FDE}=\left|Y_i^{\mathrm{p}}(T_{\mathrm{end}})-\widehat{Y}_i^{\mathrm{p}}(T_{\mathrm{end}})\right| \tag{25}$$

where N is the number of predicted timesteps, and $Y_i^{\mathrm{p}}(t)$ is the ground-truth value of location of pedestrian of interest i at time t.

Similarly, we used the average angular error (AAE) and the final angular error (FAE) to evaluate the predicted visual orientation:

$$\mathrm{AAE}=\frac{1}{N}\sum\nolimits_{t=1}^{N}\left|Y_i^{\mathrm{a}}(t)-\widehat{Y}_i^{\mathrm{a}}(t)\right| \tag{26}$$

$$\mathrm{FAE}=\left|Y_i^{\mathrm{a}}(T_{\mathrm{end}})-\widehat{Y}_i^{\mathrm{a}}(T_{\mathrm{end}})\right| \tag{27}$$

where $Y_i^{\mathrm{a}}(t)$ is the ground-truth value of visual orientation of pedestrian of interest i at time t, and $\widehat{Y}_i^{\mathrm{a}}(t)$ is the optimal predicted visual orientation.

4.1.3. Training details

In our experiments, we observed the trajectories and visual orientations of nine timesteps (3 s) and tried to predict the next $N = 9$ timesteps (3 s). The pedestrians' visual orientation a_i^t has the same form as the pedestrians' relative location, Δx_i^t. Thus, a sequence-to-sequence model can be used to predict both the pedestrians' locations and their visual orientations. We replaced the input of state-of-the-art trajectory-prediction methods with A_i for visual orientation training and prediction. All experiments were performed on the same personal computer (PC) with a NVIDIA RTX 3090 graphics processing unit (GPU).

For training the proxemics field decoder, the variety loss L_p was used:

$$L_p = \min_k \mid Y_i^p(t) - \widehat{Y}_i^p(t)^q|_2 \tag{28}$$

where $\widehat{Y}_i^p(t)^q$ is the qth predicted location of pedestrian of interest i.

We also applied the ℓ_2 loss L_a in order to measure the difference between the prediction and the ground truth of the attention field:

$$L_a = \mid Y_i^a(t) - \widehat{Y}_i^a(t)|_2^2 \tag{29}$$

4.2. Experimental discussion

4.2.1. Predicting the proxemics field

As the proxemics field represents the future location distribution of the pedestrian of interest, we evaluated our GIFNet using the accuracy of the predicted locations on the dataset. Recent studies on crowd forecasting have indicated that the short-term motion of pedestrians is highly predictable [39,40]. Here, we adopt a similar setting with a timespan $T = 3$ s and temporal resolution $R = 1/3$ s. As shown in Fig. 3(a), the ADE and FDE (i.e., the displacement error at the endpoint, shown as stars in Fig. 3(a)) of the predicted locations are used as the evaluation metrics. For each timestep, the predicted location with the highest probability is used to calculate the ADE and FDE. As illustrated in Table 1 [19,21,34,35,41–49], GIFNet outperforms the state-of-the-art learning-based trajectory-prediction methods (SoPhie [21], spatial–temporal graph attention network (STGAT) [34], social generative adversarial networks (SGAN) [35], social-spatial–temporal graph convolutional neural network (STGCNN) [19], sparse graph convolution network (SGCN) [41], etc.). Among these methods, only our GIFNet encodes all four kinds of features—that is, trajectory, visual orientation, neighbor trajectory, and group interaction state. SoPhie, STGAT, SGAN, social-STGCNN, and SGCN encode only the trajectory and integrate the information of all the surrounding neighbors with a relative-distance-dependent method. The baseline method "Linear" is a linear regressor that takes only the past trajectory as input. A more detailed ablative analysis is provided in Section 4.2.3.

For a more in-depth analysis of the neighbor and group interaction information, we divide the pedestrians into several categories (i.e., solitary pedestrians, members of an acquaintance group, members of a family group, group members without interaction, group members with non-physical interaction, and group members with physical interaction) and plot the statistical analysis results in Figs. 3(b)–(f). We use a nonparametric single-side Mann–Whitney U test to prove the statistical significance of the mean difference between the two groups of data. Figs. 3(b) and (c) illustrate the distribution of the ground truth versus the estimated forward (i.e., movement direction of the current timestep) and lateral (i.e., orthogonal to the forward direction) speeds. The prediction of GIFNet (red) shows a high consistency with the ground truth (black). The solitary pedestrians move faster than the grouped pedestrians in both directions ($p < 0.001, N = 12\,245$), and the pedestrians in the acquaintance group move faster than those in the family group in both directions ($p < 0.001, N = 12\,245$). However, grouped pedestrians with and without interactions show no significant difference, meaning that group interactions do not affect pedestrians' walking speed. In addition, the proxemics fields of solitary pedestrians are more dispersed than those of grouped pedestrians; that is, the walking direction of solitary pedestrians has higher uncertainty. These results indicate that being in a group directly affects a pedestrian's speed and walking direction.

The spatial organization of a walking pedestrian group can be measured by the angle (i.e., the inner angle between the neighbor and the forward orientation, θ_p in Fig. 3(d)) and the distance between the pedestrians in the group [26]. Figs. 3(e) and (f) illustrate the influence of interactions on the pair-wise distance and angle of pedestrian pairs within groups. The distances between pedestrian pairs with physical, non-physical, and no interaction increase significantly (both $p < 0.001$ and $N = 3668$). The angles of pedestrian pairs with physical and non-physical interactions are clustered at about 90°, while the angles of pedestrian pairs without interactions are smaller ($p < 0.001, N = 3668$) and dispersed with higher uncertainty. The distributions of pair-wise distance and angle are presented in Fig. 3(g). A total of 559 pairs of pedestrians with no interaction, 137 with non-physical interaction, and 199 with physical interaction are plotted on three 2D histograms: The pedestrian of interest is located in the center, with a 90° forward direction, and the neighbors are plotted based on distance and inner angle. Pedestrian pairs with interaction are more concentrated than those without interaction and tend to walk in parallel (i.e., angles clustered at 90°).

Fig. 3(h) illustrates the changes in the proxemics field and the pair-wise distance at the initiating time, during, and at the ending time of physical and non-physical interaction. We randomly sample 400 pairs for each state to plot the time–distance curve (bottom part of Fig. 3(h)), and the proxemics field of a representative pair is plotted in the top part of Fig. 3(h) for each stage. Pedestrian pairs with both physical and non-physical interaction show similar changes: When the pedestrians initiate interaction, they move close to each other; during the interaction, the distance between them remains stable; and, when ceasing interaction, they tend to separate. Compared with non-physical interaction, pedestrian pairs with physical interaction have smaller pair-wise distances. The high correlation between the predicted curves and the ground truth shows that GIFNet can effectively capture the changes in the group interaction state and predict accurate future locations under all states.

4.2.2. Predicting the attention field

As depicted in Fig. 1(c), the attention field is an angular range denoting the visual attention of the pedestrian of interest. Here, we fix the angular range at 30°, corresponding to the aperture of the cone of visual attention [41], and predict its central orientation. The ground-truth attention fields are calculated from the annotated visual orientations in the dataset. Similarly, we set the timespan $T = 3$ s and the temporal resolution $R = 1/3$ s, and evaluate GIFNet using the AAE and FAE. Since there is no visual orientation prediction method, we modify state-of-the-art trajectory-prediction methods for visual orientation prediction, denoted as SoPhie, STGAT, SGAN, social-STGCNN, SGCN, and so forth. "Linear" denotes the linear regression method. Table 2 [19,21,34,35,41–49] shows that our GIFNet achieves the best AAE and FAE among all the methods. As in the proxemics field prediction, the group neighbor and interaction information have a notable impact on the attention field prediction.

For a more in-depth analysis of the influence of such information on pedestrian anticipation, we evaluated the

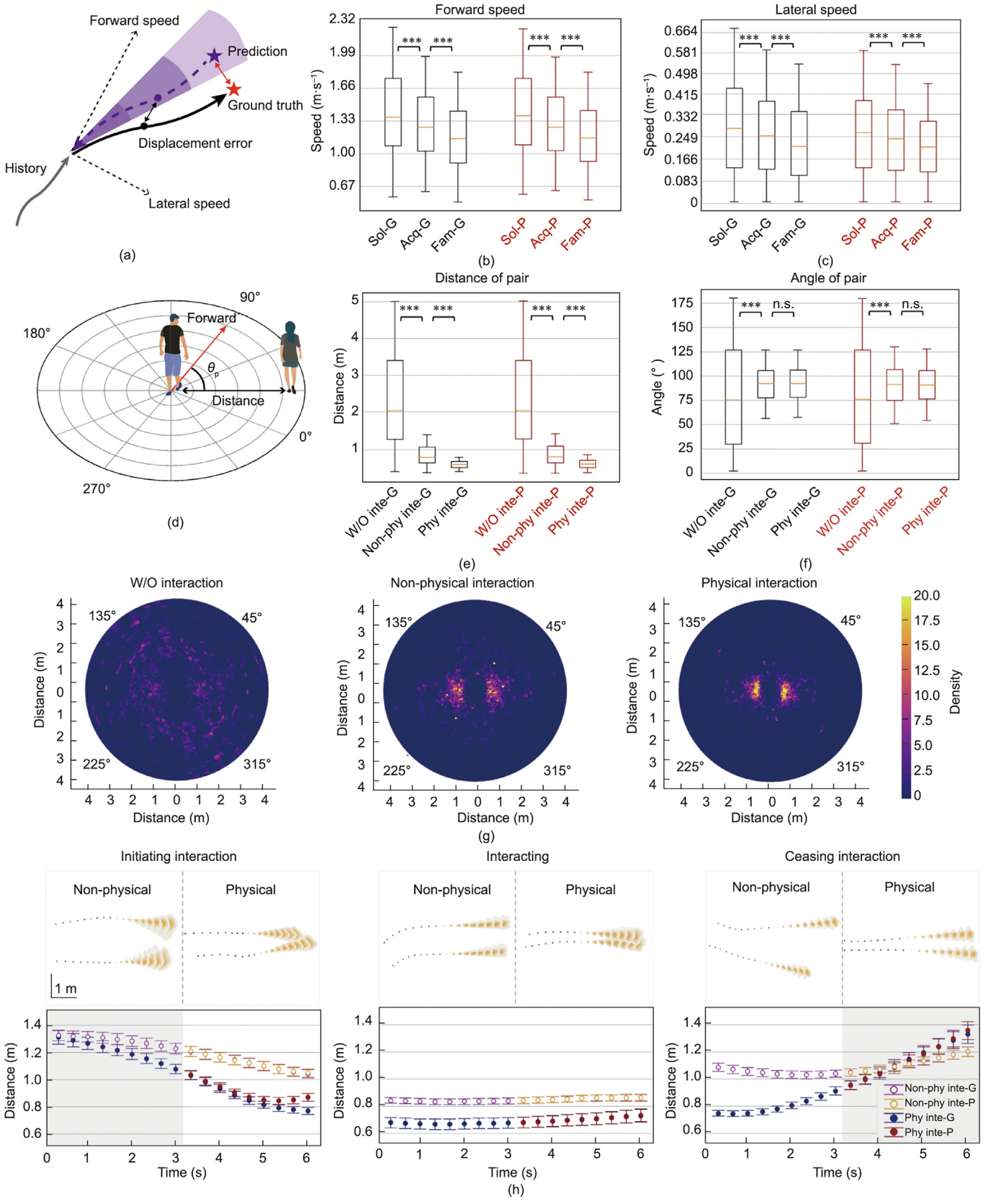
Forward speed
Prediction
Ground truth
Displacement error
History
Lateral speed
(a)
Forward speed
Speed (m·s⁻¹)
Sol-G
Acq-G
Fam-G
Sol-P
Acq-P
Fam-P
(b)
Lateral speed
(c)
90°
180°
0°
270°
Forward
Distance
(d)
Distance of pair
Distance (m)
W/O inte-G
Non-phy inte-G
Phy inte-G
W/O inte-P
Non-phy inte-P
Phy inte-P
(e)
Angle of pair
Angle (°)
n.s.
(f)
W/O interaction
Non-physical interaction
Physical interaction
135°
45°
225°
315°
Density
(g)
Initiating interaction
Interacting
Ceasing interaction
Non-physical
Physical
1 m
Time (s)
Non-phy inte-G
Non-phy inte-P
Phy inte-G
Phy inte-P
(h)

Table 1
Performance comparison of GIFNet and other state-of-the-art methods on the PANDA dataset.

Method	ADE	FDE	Ref.
SGAN	0.372	0.773	[35]
STGAT	0.392	0.840	[34]
SoPhie	0.351	0.751	[21]
Social-STGCNN	0.449	0.830	[19]
STAR	0.564	1.354	[42]
SGCN	0.353	0.718	[41]
AgentFormer	0.680	1.455	[43]
Social-Implicit	0.432	0.850	[44]
GPGraph	0.467	0.992	[45]
SocialVAE	0.408	0.908	[46]
MID	0.613	1.339	[47]
NPSN	0.485	1.039	[48]
ScePT	0.330	0.687	[49]
GIFNet	0.308	0.642	—

STGAT: spatial–temporal graph attention network; SGAN: social generative adversarial network; STGCNN: spatial–temporal graph convolutional neural network; SGCN: sparse graph convolution network. STAR: sparse trained articulated human body regressor; MID: motion indeterminacy diffusion; NPSN: non-probability sampling network.

Table 2
Performance comparison of GIFNet and other state-of-the-art methods on the PANDA dataset.

Method	AAE	FAE	Ref.
SGAN	25.523	35.982	[35]
STGAT	23.591	33.144	[34]
SoPhie	22.694	31.758	[21]
Social-STGCNN	29.091	43.793	[19]
STAR	61.603	63.569	[42]
SGCN	26.399	41.162	[41]
AgentFormer	72.788	96.895	[43]
Social-Implicit	27.901	38.795	[44]
GPGraph	26.261	43.271	[45]
SocialVAE	23.225	41.409	[46]
MID	36.724	64.442	[47]
NPSN	27.574	50.514	[48]
ScePT	18.683	32.075	[49]
GIFNet	17.447	29.735	—

forward-attention angle (θ_1 in Fig. 4(a)), cross-attention angle (θ_2 in Fig. 4(a)), and neighbor-attention angle (θ_3 in Fig. 4(a)). The forward-attention angle measures the consistency of the pedestrians' attention orientation and forward direction, the cross-attention angle measures the consistency of the attention orientation of pedestrian pairs, and the neighbor-attention angle reflects whether pedestrians' attention is attracted by their neighbors. We used a nonparametric single-side Mann-Whitney U test to demonstrate the statistical significance of the mean difference between the two groups of data. As illustrated in Figs. 4(b)–(f), all three angles predicted by our GIFNet (red) show good consistency with the ground truth (black).

As shown in Figs. 4(b), (d), and (f), the forward-attention angles of pedestrians with physical interaction, without interaction, and with non-physical interaction increase significantly (both $p < 0.001, N = 893$), the cross-attention angle of pedestrian pairs with non-physical interaction is significantly smaller than that of pedestrian pairs without interaction ($p < 0.001, N = 2986$), and the neighbor-attention angle of pedestrian pairs with physical interaction, no interaction, and non-physical interaction decreases significantly (both $p < 0.001$ and $N = 2986$). These results indicate that grouped pedestrians in physical interaction tend to focus on the direction forward (forward-attention angles close to 0°, cross-attention angles close to 0°, and neighbor-attention angles close to 90°), while pedestrians with non-physical interaction are more likely to look at each other. This may be because non-physical interactions mainly include verbal communication and eye-to-eye behaviors that require visual attention, while pedestrians in physical interaction can be more focused on walking because they can effectively perceive the location of the partner through touch instead of sight. As shown in Figs. 4(c) and (e), the forward-attention angles of the pedestrians in a family group, solitary pedestrians, and those in an acquaintance group increase significantly (both $p < 0.001$ and $N = 4641$), and the cross-attention angle of pedestrian pairs in an acquaintance group is significantly smaller than that of pedestrian pairs in the family group ($p < 0.001$, N = 12 946). This may be because family members are more likely to interact with each other physically, while acquaintances are almost equally likely to interact with each other physically and non-physically. Hence, group interactions have a significant and diverse effect on pedestrians' visual attention.

We further demonstrate the changes in the attention field at the initiating time, during, and at the ending time of physical and non-physical interactions. Similar to Fig. 3(h), the top row of Fig. 4(g) shows the representative predicted attention fields for pedestrian pairs, and the bottom row shows the changes in the neighbor-attention angle. When the pairs start to interact, the pedestrians tend to look at each other; during the interaction, both pedestrians look forward, while sometimes looking at each other (more often during non-physical interactions); when interaction ceases, the pedestrians look at each other again, and then turn to forward orientations. Similar to Fig. 3(h), the curve in Fig. 4(g) shows the high correlation between the predicted results and the ground truth, demonstrating GIFNet's ability to capture the influence of group interactions on the attention field.

4.2.3. Ablation study

We conducted a careful ablation study to demonstrate the capacity of GIFNet. As shown in Fig. 5, for the proxemics field prediction, the visual orientation information input and the group interaction information input improve the performance at every timestep. For the attention field prediction, the trajectory information input and the group interaction information input improve the performance in the first six timesteps.

From a technical perspective, the effective encoding of the group and group interactions contributes to the superior accuracy

Fig. 3. Evaluation of the proxemics field prediction. (a) Illustration of ADE and FDE. (b, c) Boxplots of the (b) forward and (c) lateral speed of solitary pedestrians and pedestrians in various groups (i.e., family and acquaintance). (d) Illustration of the angle (i.e., inner angle between the pedestrian's neighbor and forward orientation, θ_p) and distance between pedestrians. (e, f) Boxplots of the (e) pair-wise distance and (f) angle (θ_p) of grouped pedestrians without interaction, with non-physical interaction, and with physical interaction. (g) Top-view distribution maps of neighbors' location. From left to right: pedestrian pairs without interaction, with non-physical interaction, and with physical interaction. (h) Change in the proxemics field of pedestrian pairs when the group interaction state changes. The top row shows representative estimated proxemics fields with non-physical or physical interaction; blue points are input observations, and orange probability distribution maps are predicted proxemics fields. The bottom row shows the changes in the pair-wise distance. A grey background indicates no interaction, and a white background indicates that interaction is in progress. Error bars represent the standard error of the mean (SEM). Red boxplots in parts (b–f): prediction of GIFNet; black boxplots: ground truth. $^{***}p < 0.001$; n.s.: no significance, single-side Mann–Whitney U test; sol: solitary; fam: family; acq: acquaintance; inte: interaction; phy: physical; W/O: without; G: ground truth; P: prediction.

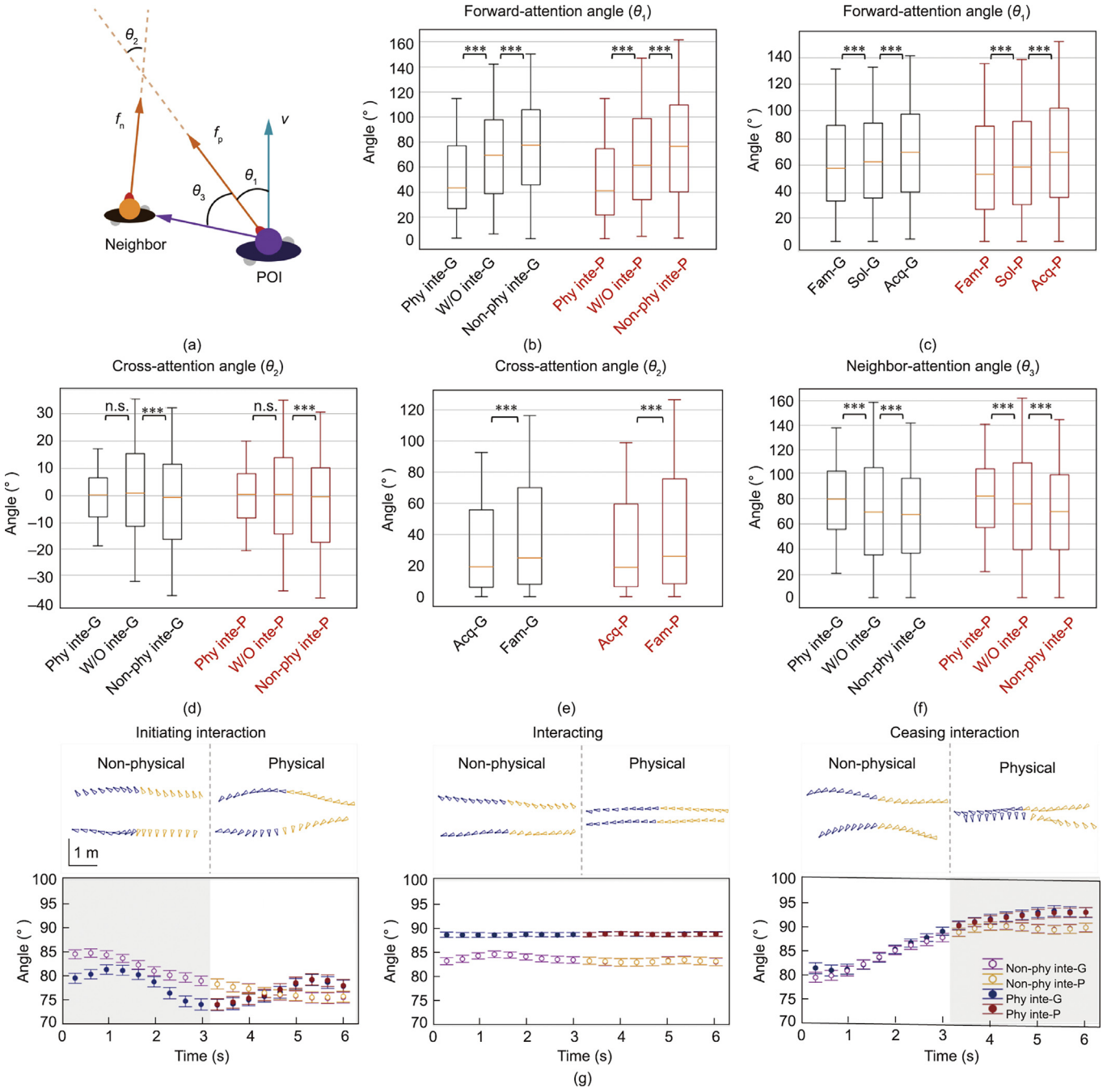

Fig. 4. Evaluation of the attention field prediction. (a) Illustration of the forward-attention angle (θ_1; the absolute value of the inner angle of visual orientation concerning the forward orientation), the cross-attention angle (θ_2; the inner angle between visual orientations of the pedestrian of interest f_p and the neighbor f_n; the angle is positive when the two orientations converge outward and negative when the two orientations converge inward), and the neighbor-attention angle (θ_3; the absolute value of the inner angle of visual orientation concerning the orientation of the neighbor). (b, c) Distributions of the ground truth (black) and predicted (red) forward-attention angle (θ_1 in part (a)) for different (b) group interaction and (c) neighbor states. (d, e) Distribution of the ground truth (black) and predicted (red) cross-attention angle (θ_2 in part (a)) for different (d) group interaction and (e) neighbor states. (f) Distribution of the ground truth and predicted neighbor-attention angles (θ_3 in part (a)). (g) Change in the attention field of pedestrian pairs when the group interaction state changes. The top row shows representative estimated attention fields with non-physical or physical interaction. Blue fans are input observations, whose center is the location of the pedestrians and whose direction is toward the attention orientation, and orange fans are predictions for future timesteps. The bottom row shows the changes in the neighbor-attention angles. A grey background indicates no interaction, and a white background indicates that interaction is in progress.

of our method. Although pooling-like operations [50] and GNNs [22,34,51–53] have been used in existing machine learning methods to encode influences among pedestrians, only the relative spatial distance between pedestrians is used in such studies, while the group and group interactions are not well considered. In addition to encoding physical features such as spatial distance, GIFNet introduces a group interaction graph with a graph attention module to propagate the group neighbor information. In this way, GIFNet explicitly reasons the importance of each group neighbor from the relative displacement and dynamic interaction states to enable better quantification of the influence of group behavior. Anthropologists recognize that visual orientation is strongly related to pedestrians' walking paths [54], and visual perception is conducive to forming group cohesion [55,56]. However, most

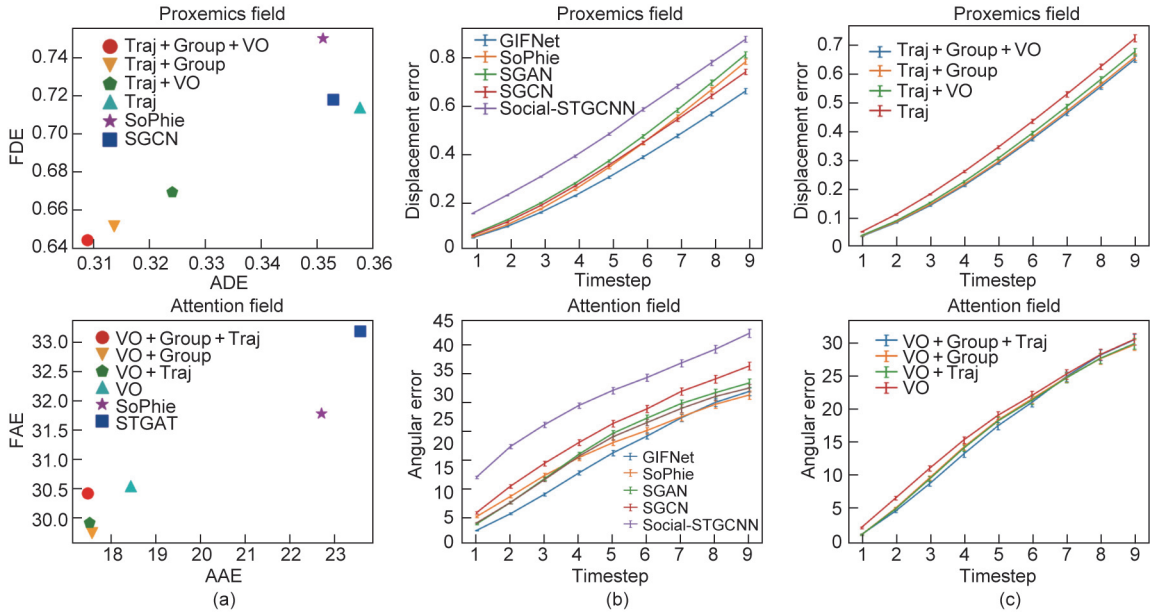

Fig. 5. An ablation analysis of the network. (a) Top: comparison of the GIFNet with full inputs including trajectories, visual orientation information, and group interaction information (red) with the inputs without visual orientation information (orange), without group interaction information (green), and without both (turquoise) when predicting the proxemics field. Two baseline methods with the best performance, SoPhie and SGCN, are also compared. Bottom: comparison of the GIFNet with full inputs including visual orientation information, group interaction information, and trajectories (red) with the inputs without trajectories (orange), without group interaction information (green), and without both (turquoise) when predicting the attention field. Two baseline methods with the best performance, SoPhie and STGAT, are also compared. (b) Comparison of the errors of GIFNet and other baseline methods on each prediction timestep (timespan $T = 3$ s and temporal resolution $R = 1/3$ s; nine timesteps in total). For proxemics field prediction, GIFNet performs best at every timestep. For attention field prediction, GIFNet performs best in the first six timesteps, but is slightly worse than SoPhie in the last three timesteps. (c) An ablative comparison of GIFNet on each prediction timestep (timespan $T = 3$ s and temporal resolution $R = 1/3$ s, nine timesteps in total). Top: comparison of the GIFNet with full inputs including trajectories, visual orientation information, and group interaction information (blue) with the inputs without visual orientation information (orange), without group interaction information (green), and without both (red) when predicting the proxemics field. Bottom: comparison of the GIFNet with full inputs including visual orientation information, group interaction information, and trajectories (blue) with the inputs without trajectories (orange), without group interaction information (green), and without both (red) when predicting the attention field. Traj: trajectory; VO: visual orientation.

of the existing methods analyze the pedestrian trajectory and visual orientation separately. In contrast, GIFNet simultaneously encodes both types of information, which mutually improves the prediction accuracy of the proxemics field and the attention field.

4.2.4. GIFNet with group detection algorithm

With the development of computer-vision-based pedestrian motion perception technology (e.g., pedestrian detection and tracking), pedestrians can be accurately positioned in video and used as input for trajectory prediction. We consider that it is also possible to realize group perception/inference; in fact, there are numerous studies on this topic, including Refs. [57,58]. Instead of taking group annotations as inputs, we supplemented a series of experiments that used the result of the group perception algorithm as input, in order to test the usability of GIFNet. We tested two methods to detect pedestrian groups: self-supervised human group detection (SHGD) [61] and correlation clustering [62]. As shown in Table 3 [57,58], compared with the baseline method that does not use group information, the application of both algorithms effectively improves GIFNet's prediction accuracy. In particular, when using the SHGD algorithm results as input, the performance is very close to that of using group annotations as input. Based on these results, we believe that many computer vision algorithms for detecting group states, tracking pedestrians, and recognizing visual orientation can be easily combined with our algorithms, which implies the usability of our algorithms in the real world. In Sections 4.2.1 and 4.2.2, given the rigor of the evaluation, we use annotations from datasets (which can be seen as "ideal" perception data) as input to our algorithms. In this way, noise from other models and several unknown uncertainties can be eliminated, allowing a more accurate assessment of our algorithm's true performance.

4.2.5. GIFNet for understanding pedestrian anticipation in a small-scale scene

Although GIFNet is designed for understanding pedestrian anticipation in large-scale scenes (in the PANDA dataset), we also evaluated how GIFNet performs in small-scale scenes (in the ETH + UCY datasets). Since the ETH + UCY datasets contain only pedestrian trajectory information, we remove the pedestrian face orientation encoder in GIFNet and use the group states detected by the SHGD algorithm as input. The ETH and UCY dataset group consists of five different scenes: ETH and Hotel (from ETH), and Univ, Zara1, and Zara2 (from UCY). Table 4 shows the ADE + FDE comparison of GIFNet and other state-of-the-art methods on the ETH + UCY datasets. The performance of GIFNet is comparable with those of SGCN and Social-Implicit in terms of the average error.

Table 3
GIFNet's performance using different group inputs.

Condition	ADE	FDE	AAE	FAE
GIFNet without group annotation	0.323	0.668	18.334	30.530
GIFNet with group annotation	0.308	0.642	17.447	29.735
GIFNet with SHGD [57]	0.310	0.648	17.473	30.317
GIFNet with correlation clustering [58]	0.318	0.662	18.173	30.466

Table 4
Performance comparison of GIFNet and other state-of-the-art methods on the ETH + UCY datasets (ADE/FDE; unit: m).

Method	ETH	Hotel	Univ	Zara1	Zara2	Average
SGAN	0.81/1.52	0.72/1.61	0.60/1.26	0.34/0.69	0.42/0.84	0.58/1.18
STGAT	0.65/1.12	0.35/0.66	0.52/1.10	0.34/0.69	0.29/0.60	0.43/0.83
SoPhie	0.70/1.43	0.76/1.67	0.54/1.24	0.30/0.63	0.38/0.78	0.54/1.15
Social-STGCNN	0.64/1.11	0.49/0.85	0.44/0.79	0.34/0.53	0.30/0.48	0.44/0.75
STAR	0.36/0.65	0.17/0.36	0.31/0.62	0.29/0.52	0.22/0.46	0.26/0.53
SGCN	0.57/1.00	0.31/0.53	0.37/0.67	0.29/0.51	0.22/0.42	0.35/0.63
AgentFormer	0.26/0.39	0.11/0.14	0.26/0.46	0.15/0.23	0.14/0.24	0.18/0.29
Social-Implicit	0.66/1.44	0.20/0.36	0.31/0.60	0.25/0.50	0.22/0.43	0.33/0.67
GPGraph	0.43/0.63	0.18/0.30	0.24/0.42	0.17/0.31	0.15/0.29	0.23/0.39
SocialVAE	0.41/0.58	0.13/0.19	0.21/0.36	0.17/0.29	0.13/0.22	0.21/0.33
MID	0.39/0.66	0.13/0.22	0.22/0.45	0.17/0.30	0.13/0.27	0.21/0.38
NPSN	0.36/0.68	0.11/0.18	0.18/0.39	0.14/0.29	0.11/0.23	0.18/0.35
ScePT	0.10/0.65	0.13/0.77	0.12/0.65	0.13/0.77	0.14/0.81	0.12/0.73
GIFNet	0.36/0.74	0.38/0.74	0.30/0.61	0.40/0.87	0.34/0.76	0.36/0.74

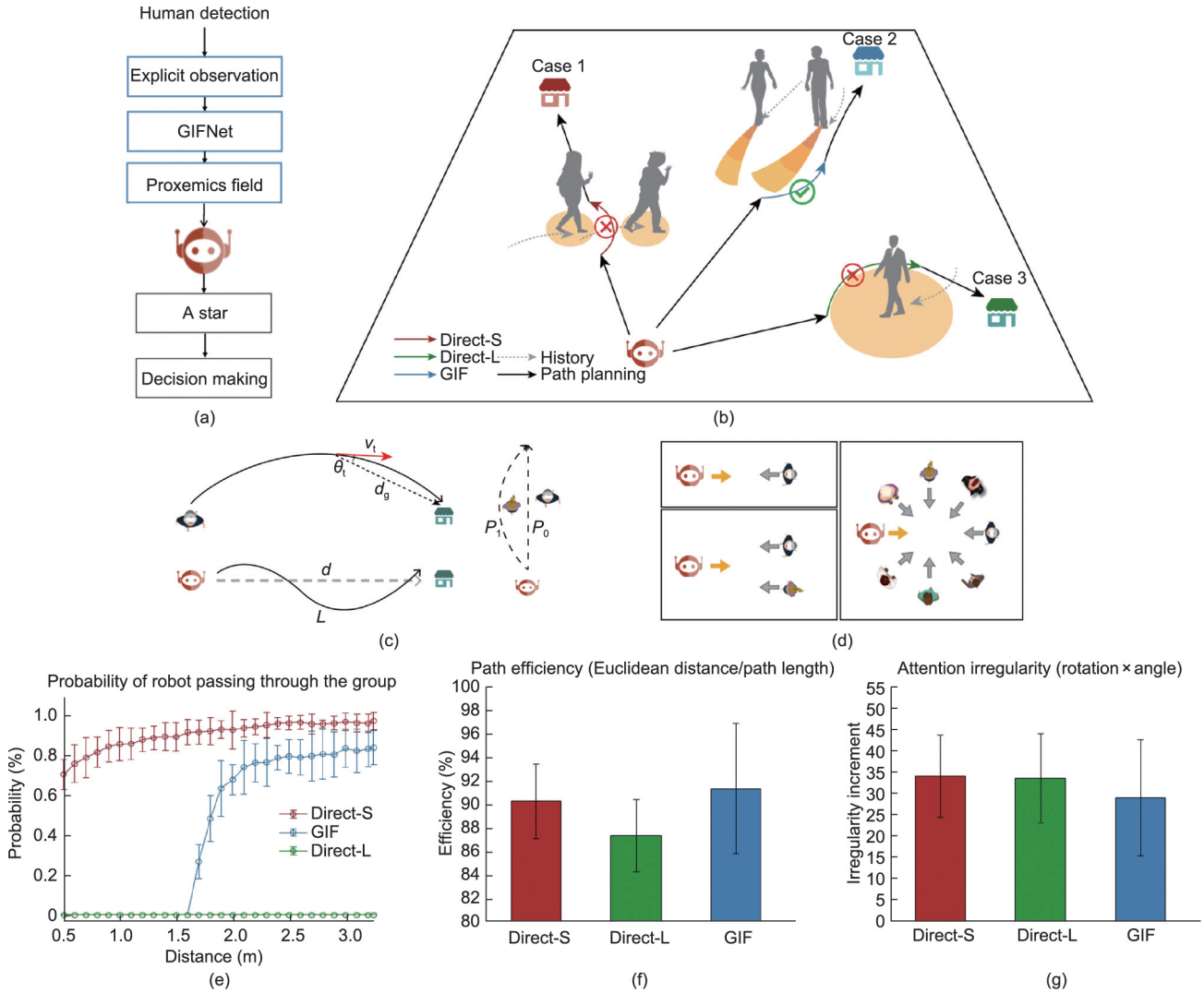

Fig. 6. The GIF for crowd-aware robot navigation. (a) Illustration of the pipeline of GIF-based robot navigation. Our method uses the GIF to guide robot path planning. (b) Illustration of three different path-planning methods in our experiments. The Direct-S and Direct-L methods are widely used robot navigation approaches that treat pedestrians as simple circular obstacles [60]. The Direct-S method treats pedestrians as small circular obstacles and often causes robots to disturb pedestrians. The Direct-L method treats pedestrians as large circular obstacles, resulting in sacrifices in navigation efficiency and a longer path. The GIF method treats the range of the predicted proxemics field as an obstacle. To ensure a fair comparison, we adopt the classic A-star path-planning method to evaluate the performance of the three methods. (c) Three metrics are used for evaluation. The robot path efficiency is the ratio between the robot's Euclidean distance d to the goal and the actual traveled distance L, which measures the robot navigation efficiency. The attention irregularity is the sum of the pedestrians' unnecessary rotation angle caused by the robot, which measures the robot's disturbance to pedestrians. The probability of the robot passing through the group P_0 is used to measure the disturbance the robot causes to the pedestrian group. (d) Illustration of the layout of our experiments. We evaluate three methods under three different scenarios: one robot with one person, one robot with two persons, and one robot with multiple persons. (e) The probability of the robot passing through a two-person group versus the inner distance between the two people in the two-person group. Our group-aware method can accurately identify groups (whose inner distance is usually less than 1.5 m) and prevent robots from passing through them. (f, g) A comparison of the (f) path efficiency and (g) attention irregularity. Our GIF-based method has the highest path efficiency and the lowest attention irregularity, which suggests that our method effectively prevents robots from disturbing pedestrians while maintaining the robot's driving efficiency. The error bars represent the standard deviation. θ_t: the pedestrians' rotation angle; d_g: the direction of destination; v_t: the direction of current speed.

Due to a lack of facial orientation information and the use of algorithmic inferred groups, GIFNet does not take full advantage of the benefits provided by its innovative structural design. However, the above results are sufficient to demonstrate that GIFNet is an advanced pedestrian trajectory-prediction algorithm that is applicable to different datasets.

4.2.6. The GIF for crowd-aware robot navigation

With the booming development of unmanned systems (e.g., autonomous driving, service robots, etc.), such systems' environments are envisioned to expand from isolated areas to social spaces shared with humans. People expect unmanned systems to not only have powerful functions but also provide smart interactions with comfort, naturalness, and sociability [59]. Our proposed group-aware understanding of pedestrian anticipation may enable unmanned systems to work in a human-like manner and comply with social norms, which is shown in Fig. 6. As a validation, we propose a new robot navigation method based on the GIF (Fig. 6(a)). Existing robot navigation approaches usually regard pedestrians as simple circular obstacles and avoid pedestrians based on their current locations [60], which makes it difficult to strike a balance between not disturbing pedestrians and maintaining navigation efficiency. In contrast, by imparting the robot with the human-like capability of anticipation, our method can adaptively plan the robot's path according to the pedestrians' proxemics field and attention field, which reflect the pedestrians' behavioral intention in a fine-grained way, thus effectively preventing the robot from disturbing pedestrians while maintaining the robot's driving efficiency (Figs. 6(e)–(g)). We believe that the GIF can promote a harmonious human–machine relationship in broader applications.

5. Conclusions

Understanding pedestrian anticipation is a long-standing problem with significant application value. In this paper, we studied how different group relationships influence pedestrian anticipation. More specifically, we proposed the GIF, a novel group-aware representation of pedestrian anticipation, which can quantitatively explain how people's anticipation of others' speed, others' attention, and the spatial organization of groups is dynamically affected by group interactions. Furthermore, we tailored GIFNet to estimate the GIF based on the explicit observations of pedestrians. By encoding multidimensional data, including pedestrian trajectory, visual attention, and state of group interaction, GIFNet succeeds in representing changes in pedestrian anticipation under the prominent impact of group behaviors and in accurately predicting the future states of pedestrians.

In practice, the GIF will contribute to a group-aware-based understanding of pedestrian anticipation and pedestrian group behavior. The GIF can enable unmanned systems to accurately anticipate pedestrians' actions and safely and comfortably interact with them, thereby promoting a harmonious human–machine relationship. We believe that the GIF has enormous potential for application in interdisciplinary areas, such as intelligent unmanned systems, the social-aware modeling of pedestrian dynamics, and emergency evacuation.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC; 62125106, 61860206003, and 62088102), in part by the Ministry of Science and Technology of China (2021ZD0109901), and in part by the Provincial Key Research and Development Program of Zhejiang (2021C01016).

Compliance with ethics guidelines

Xueyang Wang, Xuecheng Chen, Puhua Jiang, Haozhe Lin, Xiaoyun Yuan, Mengqi Ji, Yuchen Guo, Ruqi Huang, and Lu Fang declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Rasouli A, Kotseruba I, Tsotsos JK. Pedestrian action anticipation using contextual feature fusion in stacked RNNs. In: Proceedings of the 30th British Machine Vision Conference (BMVC 2019); 2019 Sep 9–12; Cardiff, UK. London: BMVA Press; 2019.
[2] Luo Y, Cai P, Bera A, Hsu D, Lee WS, Manocha D. PORCA: modeling and planning for autonomous driving among many pedestrians. IEEE Robot Autom Lett 2018;3(4):3418–25.
[3] Trautman P, Ma J, Murray RM, Krause A. Robot navigation in dense human crowds: the case for cooperation. In: Proceedings of the IEEE International Conference on Robotics and Automation; 2013 May 6–10; Karlsruhe, Germany. New York City: IEEE; 2013. p. 2153–60.
[4] Yao X, Zhang J, Oh J. Following social groups: socially compliant autonomous navigation in dense crowds. 2019. arXiv:1911.12063.
[5] Zhou J, Shi ZK. A new lattice hydrodynamic model for bidirectional pedestrian flow with the consideration of pedestrian's anticipation effect. Nonlinear Dyn 2015;81(3):1247–62.
[6] Hoogendoorn S, Bovy PHL. Simulation of pedestrian flows by optimal control and differential games. Optim Control Appl Methods 2003;24:153–72.
[7] Zheng X, Cheng Y. Conflict game in evacuation process: a study combining Cellular Automata model. Physica A Stat Mech Appl 2011;390:1042–50.
[8] Bouzat S, Kuperman MN. Game theory in models of pedestrian room evacuation. Phys Rev E Stat Nonlinear Soft Matter Phys 2014;89:032806.
[9] Xu Q, Chraibi M, Seyfried A. Anticipation in a velocity-based model for pedestrian dynamics. Transp Res Part C Emerg Technol 2021;133:103464.
[10] Suma Y, Yanagisawa D, Nishinari K. Anticipation effect in pedestrian dynamics: modeling and experiments. Physica A Stat Mech Appl 2012;391:248–63.
[11] Nowak S, Schadschneider A. Quantitative analysis of pedestrian counterflow in a cellular automaton model. Phys Rev E Stat Nonlin Soft Matter Phys 2012;85(6):066128.
[12] Bailo R, Carrillo JA, Degond P. Pedestrian models based on rational behaviour. In: Gibelli L, Bellomo N, editors. Crowd dynamics. Volume 1—modeling and simulation in science, engineering and technology. Berlin: Springer; 2018.
[13] Murakami H, Feliciani C, Nishiyama Y, Nishinari K. Mutual anticipation can contribute to self-organization in human crowds. Sci Adv 2021;7(12):eabe7758.
[14] Murakami H, Feliciani C, Nishinari K. Lévy walk process in self-organization of pedestrian crowds. J R Soc Interface 2019;16(153):20180939.
[15] Roe RM, Busemeyer JR, Townsend JT. Multialternative decision field theory: a dynamic connectionist model of decision making. Psychol Rev 2001;108(2):370–92.
[16] Karamouzas I, Skinner B, Guy SJ. Universal power law governing pedestrian interactions. Phys Rev Lett 2014;113(23):238701.
[17] Zanlungo F, Ikeda T, Kanda T. Social force model with explicit collision prediction. EPL 2011;93(6):68005.
[18] Kosaraju V, Sadeghian A, Martín-Martín R, Reid I, Rezatofighi H, Savarese S. Social-BiGAT: multimodal trajectory forecasting using bicycle-GAN and graph attention networks. In: Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019); 2019 Dec 8–14; Vancouver, BC, Canada; 2019.
[19] Mohamed A, Qian K, Elhoseiny M, Claudel C. Social-STGCNN: a social spatio–temporal graph convolutional neural network for human trajectory prediction. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 14–19; online. New York City: IEEE; 2020. p. 14424–32.
[20] Rudenko A, Palmieri L, Lilienthal AJ, Arras KO. Human motion prediction under social grouping constraints. In: Proceedings of the IEEE International Workshop on Intelligent Robots and Systems (IROS 2018); 2018 Oct 1–5; Madrid, Spain. New York City: IEEE; 2018. p. 3358–64.
[21] Sadeghian A, Kosaraju V, Sadeghian A, Hirose N, Rezatofighi SH, Savarese S. SoPhie: an attentive GAN for predicting paths compliant to social and physical constraints. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019 Jun 16–20; Long Beach, CA, USA. New York City: IEEE; 2019. p. 1349–58.
[22] Sun J, Jiang Q, Lu C. Recursive social behavior graph for trajectory prediction. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019 Jun 16–20; Long Beach, CA, USA. New York City: IEEE; 2019. p. 660–9.
[23] Mangalam K, Girase H, Agarwal S, Lee KH, Adeli E, Malik J, et al. It is not the journey but the destination: endpoint conditioned trajectory prediction. In: Proceedings of the 2020 European Conference on Computer Vision; 2020 Aug 23–28; Glasgow, UK. Berlin: Springer; 2020. p. 759–76.
[24] Salzmann T, Ivanovic B, Chakravarty P, Pavone M. Trajectron++: dynamically-feasible trajectory forecasting with heterogeneous data. In: Proceedings of the

2020 European Conference on Computer Vision; 2020 Aug 23–28; Glasgow, UK. Berlin: Springer. p. 683–700.
[25] Zhou C, Han M, Liang Q, Hu YF, Kuai SG. A social interaction field model accurately identifies static and dynamic social groupings. Nat Hum Behav 2019;3(8):847–55.
[26] Moussaïd M, Perozo N, Garnier S, Helbing D, Theraulaz G. The walking behaviour of pedestrian social groups and its impact on crowd dynamics. PLoS One 2010;5(4):e10047.
[27] Liu Y, Yan Q, Alahi A. Social NCE: contrastive learning of socially-aware motion representations. In: Proceedings of the 2020 IEEE/CVF International Conference on Computer Vision; 2020 Jun 13–19; Seattle, WA, USA. New York City: IEEE; 2020. p. 15118–29.
[28] De Jaegher H, Di Paolo E, Gallagher S. Can social interaction constitute social cognition? Trends Cogn Sci 2010;14(10):441–7.
[29] Cheng L, Yarlagadda R, Fookes CB, Yarlagadda PK. A review of pedestrian group dynamics and methodologies in modelling pedestrian group behaviours. World J Mech Eng 2014;1:1–13.
[30] Yücel Z, Zanlungo F, Shiomi M. Modeling the impact of interaction on pedestrian group motion. Adv Robot 2018;32(3):137–47.
[31] Zhou R, Zhou H, Gao H, Tomizuka M, Li J, Xu Z. Grouptron: dynamic multi-scale graph convolutional networks for group-aware dense crowd trajectory forecasting. In: Proceedings of the 2022 International Conference on Robotics and Automation (ICRA 2022); 2022 May 23–27; Philadelphia, PA, USA. New York City: IEEE; 2020. p. 805–11.
[32] Casas S, Gulino C, Liao R, Urtasun R. SpAGNN: spatially-aware graph neural networks for relational behavior forecasting from sensor data. In: Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA 2020); 2020 May 31–Aug 31; online. New York City: IEEE; 2020. p. 9491–7.
[33] Girase H, Gang H, Malla S, Li J, Kanehara A, Mangalam K, et al. LOKI: long term and key intentions for trajectory prediction. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision; 2021 Oct 11–17; Montreal, BC, Canada. New York City: IEEE; 2021. p. 9803–12.
[34] Huang Y, Bi H, Li Z, Mao T, Wang Z. STGAT: modeling spatial–temporal interactions for human trajectory prediction. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. New York City: IEEE; 2019. p. 6272–81.
[35] Gupta A, Johnson J, Li FF, Savarese S, Alahi A. Social GAN: socially acceptable trajectories with generative adversarial networks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2018 Jun 18–23; Salt Lake City, UT, USA. New York City: IEEE; 2018. p. 2255–64.
[36] Zhang B, Chen W, Ma X, Qiu P, Liu F. Experimental study on pedestrian behavior in a mixed crowd of individuals and groups. Physica A Stat Mech Appl 2020;556:124814.
[37] Gallup AC, Hale JJ, Sumpter DJ, Garnier S, Kacelnik A, Krebs JR, et al. Visual attention and the acquisition of information in human crowds. Proc Natl Acad Sci USA 2012;109(19):7245–50.
[38] Wang X, Zhang X, Zhu Y, Guo Y, Yuan X, Xiang L, et al. PANDA: a gigapixel-level human-centric video dataset. In: Proceedings of the 2020 IEEE/CVF conference on computer vision and pattern recognition; 2020 Jun 14–19; online. New York City: IEEE; 2020. p. 3268–78.
[39] Raksincharoensak P, Hasegawa T, Nagai M. Motion planning and control of autonomous driving intelligence system based on risk potential optimization framework. Int J Automot Eng 2016;7(AVEC14):53–60.
[40] Alahi A, Ramanathan V, Li FF. Socially-aware large-scale crowd forecasting. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition; 2014 Jun 23–28; Columbus, OH, USA. New York City: IEEE; 2014. p. 2211–8.
[41] Shi L, Wang L, Long C, Zhou S, Zhou M, Niu Z, et al. SGCN: sparse graph convolution network for pedestrian trajectory prediction. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 19–25; online. New York City: IEEE; 2021. p. 8994–9003.
[42] Osman AAA, Bolkart T, Black MJ. STAR: sparse trained articulated human body regressor. In: Proceedings of the Computer Vision–ECCV 2020: 16th European Conference; 2020 Aug 23–28; Glasgow, UK. Berlin: Springer International Publishing; 2020. p. 598–613.
[43] Yuan Y, Weng X, Ou Y, Kitani K. AgentFormer: agent–aware transformers for socio-temporal multi-agent forecasting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021 Oct 10–17; Montreal, QC, Canada. New York City: IEEE; 2021. p. 9813–23.
[44] Mohamed A, Zhu D, Vu W, Elhoseiny M, Claudel C. Social-Implicit: rethinking trajectory prediction evaluation and the effectiveness of implicit maximum likelihood estimation. In: Proceedings of the Computer Vision–ECCV 2022: 17th European Conference; 2022 Oct 23–27; Tel Aviv, Israel. Berlin: Springer; 2022. p. 463–79.
[45] Bae I, Park JH, Jeon HG. Learning pedestrian group representations for multi-modal trajectory prediction. In: Proceedings of the Computer Vision–ECCV 2022: 17th European Conference; 2022 Oct 23–27; Tel Aviv, Israel. Berlin: Springer; 2022.
[46] Xu P, Hayet JB, Karamouzas I. SocialVAE: human trajectory prediction using timewise latents. In: Proceedings of the Computer Vision–ECCV 2022: 17th European Conference; 2022 Oct 23–27; Tel Aviv, Israel. Berlin: Springer; 2022. p. 511–28.
[47] Gu T, Chen GY, Li J, Lin C, Rao Y, Zhou J, et al. Stochastic trajectory prediction via motion indeterminacy diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 21–24; New Orleans, LU, USA. New York City: IEEE; 2022.
[48] Bae I, Park JH, Jeon HG. Non-probability sampling network for stochastic human trajectory prediction. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 21–24; New Orleans, LU, USA. New York City: IEEE; 2022.
[49] Chen Y, Ivanovic B, Pavone M. ScePT: scene-consistent, policy-based trajectory predictions for planning. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 21–24; New Orleans, LU, USA. New York City: IEEE; 2022.
[50] Kothari P, Kreiss S, Alahi A. Human trajectory forecasting in crowds: a deep learning perspective. IEEE Trans Intell Transp Syst 2021;23(7):7386–400.
[51] Yu C, Ma X, Ren J, Zhao H, Yi S. Spatio–temporal graph transformer networks for pedestrian trajectory prediction. In: Proceedings of the 2020 European Conference on Computer Vision; 2020 Aug 23–28; online. Berlin: Springer; 2020. p. 507–23.
[52] Qiu J, Tang J, Ma H, Dong Y, Wang K, Tang J. DeepInf: social influence prediction with deep learning. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2018 Aug 19–23; London, UK. New York City: Association for Computing Machinery (ACM); 2018. p. 2110–9.
[53] Liu C, Chen Y, Liu M, Shi BE. AVGCN: trajectory prediction using graph convolutional networks guided by human attention. In: Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA); 2021 May 30–Jun 5; Xi'an, China. New York City: IEEE; 2021. p. 14234–40.
[54] Hasan I, Setti F, Tsesmelis T, Del Bue A, Cristani M, Galasso F. "Seeing is believing": pedestrian trajectory forecasting using visual frustum of attention. In: Proceedings of the 2018 IEEE Workshop on Applications of Computer Vision (WACV 2018); 2018 Mar 12–15; Lake Tahoe, NV, USA. New York City: IEEE; 2018. p. 1178–85.
[55] Bastien R, Romanczuk P. A model of collective behavior based purely on vision. Sci Adv 2020;6(6):eaay0792.
[56] Lavergne FA, Wendehenne H, Bäuerle T, Bechinger C. Group formation and cohesion of active particles with visual perception–dependent motility. Science 2019;364(80):70–4.
[57] Li J, Han R, Yan H, Qian Z, Feng W, Wang S. Self-supervised social relation representation for human group detection. In: Proceedings of the Computer Vision–ECCV 2022: 17th European Conference; 2022 Oct 23–27; Tel Aviv, Israel. Berlin: Springer; 2022.
[58] Solera F, Calderara S, Cucchiara R. Socially constrained structural learning for groups detection in crowd. IEEE Trans Pattern Anal Mach Intell 2016;38(5):995–1008.
[59] Kruse T, Pandey AK, Alami R, Kirsch A. Human–aware robot navigation: a survey. Robot Auton Syst 2013;61(12):1726–43.
[60] Gul F, Rahiman W, Nazli Alhady SS, Chen K. A comprehensive study for robot navigation techniques. Cogent Eng 2019;6(1):1632046.