Causal Discovery and Deep Learning-based Interaction-aware Pedestrian Trajectory Prediction

Wen-Xin Qiu[1], Takashi Fuse[2]

Dept. of Civil Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 JAPAN
[1]qiu@civil.t.u-tokyo.ac.jp [2]fuse@civil.t.u-tokyo.ac.jp

Keywords: Causal Graph, Machine Learning, Graph Neural Network, Microscopic Pedestrian Behavior, Spatial-Temporal Data

Abstract

Understanding pedestrian behaviors is the foundation of simulation for space planning. However, conventional behavior modeling methods are insufficient for learning detailed interactions, and deep learning methods often lack interpretability. This study aims to develop a pedestrian trajectory modeling approach based on discovering causal relationships among pedestrians. The proposed method consists of two parts: analyzing causal relationships among pedestrians using statistical causal discovery methods and predicting trajectories using attention-based deep learning methods. The first part employs a semi-parametric method to identify the causal relationships underlying observed pedestrian behavior and construct a spatial-temporal graph based on these causal relationships. The second part primarily uses the graph attention network to learn interactions among pedestrians. The experimental results demonstrate that the proposed method achieves a good balance between prediction accuracy and interpretability, while also identifying limitations, including at low-density scenes and due to causal model assumptions.

1. Introduction

Understanding pedestrian behavior is essential for planning safe and comfortable pedestrian spaces. It forms the foundation for pedestrian simulations (Helbing et al., 2002; Daamen, 2004) and safety analyses (Papadimitriou et al., 2009). Moreover, autonomous mobile robots rely on pedestrian behavior models to interact smoothly with humans (Unhelkar et al., 2015); autonomous vehicles also require precise predictions of human movement for safe navigation (Brouwer et al., 2016).

Pedestrian behaviors can be categorized into three hierarchical decision-making levels: strategic, tactical, and operational levels (Hoogendoorn and Bovy, 2002; Daamen, 2004). The operational-level behavior models the instantaneous decisions made during walking, thereby shaping pedestrians' actual trajectories. Representative conventional approaches include the social force model (Helbing et al., 2002), discrete choice model-based approaches (Antonini et al., 2006; Robin et al., 2009), cellular automaton approaches (Blue and Adler, 1998; Schadschneider, 2002), and potential field approaches (Karamouzas et al., 2009). Most approaches have emphasized that pedestrians are affected by others. However, these conventional methods rely on known behavioral patterns, such as avoiding or following other agents, so they cannot model complex interaction dynamics and have limited expressive power.

Recently, pedestrian trajectory prediction approaches leveraging deep learning (DL)-based methods have been widely proposed because they can capture more detailed patterns. Besides, these approaches predict future trajectories without specifying a destination. These approaches can be classified into “pooling-based” and “attention-based” approaches, which consider either the overall situation of all pedestrians or pairwise interactions between pedestrians (Minoura et al., 2022). Social LSTM (Alahi et al., 2016) proposed the first social pooling mechanism that considers pedestrian interactions, achieved more accurate predictions than purely using LSTM (Long Short-Term Memory) layers, and has significantly advanced research on DL-based pedestrian trajectory prediction (Gupta et al. (2018); Sun et al. (2020), *etc.*). However, the social pooling mechanism cannot distinguish the effect of different individuals. For more precise interaction modeling, attention-based approaches are proposed and often combined with spatial-temporal graphs that represent pedestrian trajectories (Vemula et al. (2018); Huang et al. (2019), *etc.*). The spatial-temporal graph consists of spatial graphs representing pedestrians' interactions at each time step and temporal graphs representing each pedestrian's trajectory. The attention mechanism was initially proposed for sequence data to learn where to pay attention to (Vaswani et al., 2017). It has operated robustly with LSTM and has been dedicated to graph data (Veličković et al., 2017).

While DL-based approaches become more accurate in prediction, they lack the ability to interpret how pedestrians behave in response to others' movements. The incomprehensibility precludes their understanding of pedestrian behavior. Interpretation methods are thus proposed for DL models. For example, Kothari et al. (2021) has developed an interpretable conventional framework that integrates a discrete choice model-based approach with a DL model. On the other hand, post hoc methods, such as the Shapley value, can be used to assess the importance of input factors in black-box models. Although these methods have been used to conduct interpretations, merely describing the correlations between inputs and outputs is insufficient to understand the mechanism of pedestrian behavior. Thus, this study aims to develop an approach based on the analysis of causal relationships in pedestrian behaviors. Statistical causal discovery (CD) methods are adopted to identify variables as causes and outcomes from observed data.

The objective of this study is to predict future trajectories using a causal, quantitative understanding of pedestrian interactions. The proposed approach consists of two parts. The first part used a semi-parametric CD method to identify causal relationships among pedestrians. The CD method provides causal interpretations of behaviors rather than relying on handcrafted features or correlations. The second part utilized a graph attention-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-4-2026-275-2026

based DL model for trajectory prediction under the restriction of causal relationships. Graph attention networks (GATs) are efficient in learning and provide visualized quantitative values of interactions. The balance between accuracy and interpretability is valued.

2. Related Work

2.1 Deep Learning-based Pedestrian Trajectory Prediction

The first social pooling method, Social LSTM (Alahi et al., 2016), used a pooling layer to summarize the discretized pedestrian locations at each time step and pass these features to all pedestrians for prediction. The Social GAN (Gupta et al., 2018) eliminated coordinate discretization and adopted a Generative Adversarial Network (GAN) structure to model the multimodal nature of possible trajectories. Reciprocal Net (Sun et al., 2020) extended the social pooling structure to set the threshold of maximum considered pedestrians. These studies focused on developing detailed social pooling modules to improve prediction accuracy, but were still limited by the nature of pooling, which summarizes the effects of different pedestrians into a single piece of information.

The Social Attention (Vemula et al., 2018) adopted an attention mechanism that learns interactions between each pair of pedestrians and a spatial-temporal graph to represent these relationships, inspired by Structural RNN (Jain et al., 2016). RNNs (Recursive Neural Networks) on the edges and nodes updated the features of themselves, and attention modules between the RNNs learned the importance of pedestrian pairs. However, the recursive computation in RNNs increased model computational cost and led to the vanishing gradient problem.

Otherwise, GNNs (Graph Neural Networks) are a more efficient and effective choice for graphs. Current attention-based models typically utilize a spatial-temporal graph with GATs (Huang et al., 2019) or Transformers (Yu et al., 2020) to learn explicit representations for each pair of pedestrians. For example, Mohamed et al. (2020) and Shi et al. (2021) used graph convolutional networks (GCNs), which compute the graph Laplacian in the spectral domain, to learn graph features. Xu et al. (2018) directly calculated the attention values between the encoded motion and location information, and Sadeghian et al. (2019) also designed the physical attention and social attention modules without a graph.

2.2 Causal Discovery

Causal discovery (CD) is a category of methods that identifies causal relationships between variables. Unlike correlation, causation is a directional relationship in which outcomes cannot affect their causes. A causal relationship can be described as a causal graph, which is a directed graph with edges representing causation, or a set of functions that represent causation. General assumptions include the directed acyclic graph (DAG) and the non-existence of unobserved confounders.

CD methods can be categorized as parametric, semi-parametric, and non-parametric. The non-parametric methods, such as the Fast Causal Inference (FCI) method (Spirtes et al., 2000), can estimate a causal graph with direction of causes without making assumptions about either functions or error variables. The parametric methods (e.g., Malinsky and Spirtes (2016)), assuming linear functions and Gaussian-distributed errors, can estimate the causal graph (direction) and the causal effect (weight). However, both methods sometimes lack identifiability, meaning the causal direction of some edges cannot always be identified. The semi-parametric CD methods, represented by the LiNGAM (Linear Non-Gaussian Acyclic Model) (Shimizu et al., 2006) method, assume linear functions and non-Gaussian error distributions. The assumption of non-Gaussian distributions ensures identifiability by excluding Gaussian distributions, which remain Gaussian under linear transformations.

The concept of causality is employed in interpreting DL models. Makansi et al. (2021) introduced a Shapley value-based approach to interpret various trajectory prediction methods by analyzing the relationship between input attributes and model accuracy. However, their concept, which originated from Granger causality, does not represent strict causality. In a different context, Sani et al. (2023) investigated the characteristics of input features that influence the outcome of a classification problem. Nevertheless, this approach provides a post-hoc interpretation of the relationships between inputs and outputs, rather than offering a direct explanation of the black-box model itself.

3. Method

3.1 Problem Formulation

The pedestrian trajectory problem to be solved in this study is illustrated in Figure 1. The goal is to predict future trajectories from observed pedestrian trajectories over a fixed time window. To focus on modeling pedestrian interaction, this study excluded all environmental context from the input. Although some previous works have considered various environmental factors through inputting scene images or videos (*e.g.*, Sadeghian et al. (2019); Kosaraju et al. (2019); Shafiee et al. (2021)), the inputs for this study are only trajectories of multiple pedestrians in a short period. The inputs are mathematically defined as a set of time series of locations $\mathcal{P}_{\text{obs}} = \{p_n^t | n \in \{1, ..., N\}, t \in \{1, ..., T_{\text{obs}}\}\}$, where N is the total number of pedestrians indexed by n, and T_{obs} is the discretized observation time step indexed by t. The location of pedestrian n at time step t is denoted by p_n^t, which lies in the two-dimensional ground coordinates (x_n^t, y_n^t). The expected output is the predicted trajectories of N pedestrians in a short period of time, T_{obs}, defined as $\hat{\mathcal{P}_{\text{pred}}} = \{\hat{p_n^t} | n \in \{1, ..., N\}, t \in \{T_{\text{obs}}+1, ..., T_{\text{obs}}+T_{\text{pred}}\}\}$ and the corresponding ground truth of trajectories are defined as $\mathcal{P}_{\text{pred}} = \{p_n^t | n \in \{1, ..., N\}, t \in \{T_{\text{obs}} + 1, ..., T_{\text{obs}} + T_{\text{pred}}\}\}$. Thus, the objective of this prediction problem can be summarized as finding a model f that performs $\hat{\mathcal{P}_{\text{pred}}} = f(\mathcal{P}_{\text{obs}})$.

3.2 Proposed Framework

The framework of this study is shown in Figure 2. The proposed method consists of two parts: (1) analyzing causal relationships among pedestrians using a CD method and (2) predicting trajectories using attention-based DL methods. In the first part, the trajectories first undergo a preprocessing step. Then, CD is performed to identify causal relationships underlying observed pedestrian behavior, and a spatial-temporal (ST) graph is built based on these relationships. In the second part, the ST graph serves as input to the DL model, and the model's outputs are predictions of multivariate Gaussian probability distributions for future trajectories.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-4-2026-275-2026 | © Author(s) 2026. CC BY 4.0 License.

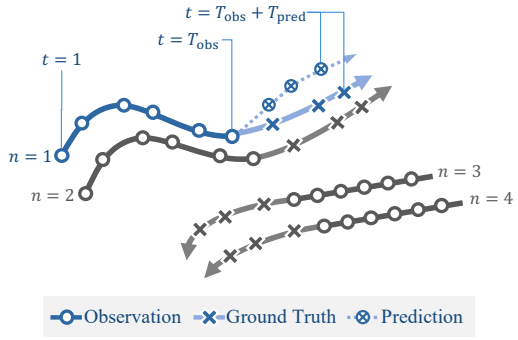


Figure 1. Illustration of pedestrian trajectory prediction with four pedestrians

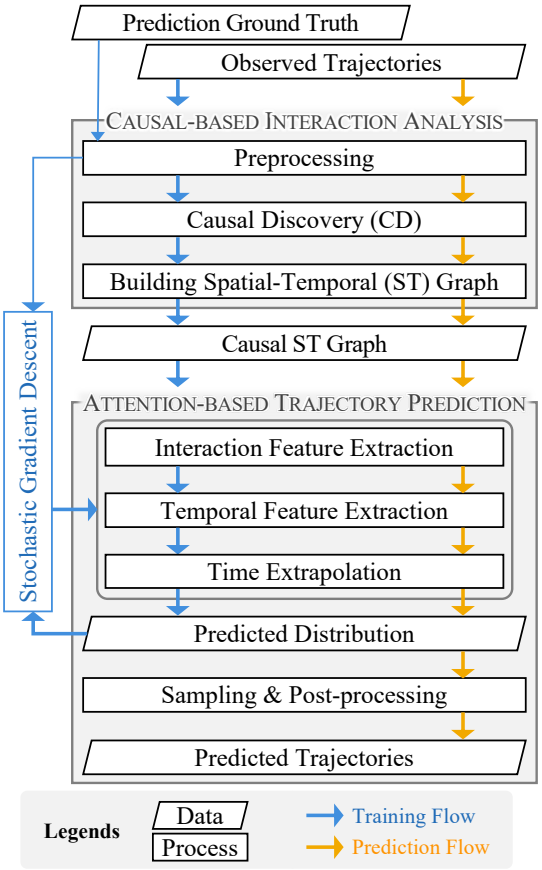


Figure 2. Framework of the proposed method

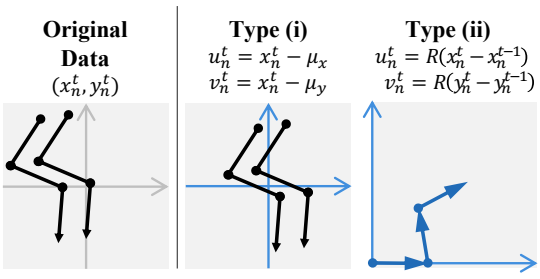


Figure 3. Two preprocessing methods

3.3 Data Preprocessing

The preprocessing step is necessary because of the large variance caused by varying coordinate settings. Additionally, normalizing the data to a limited range helps prevent gradient explosion or vanishing problems in DL models. In this study, the observed data are assumed to be recorded in real-world coordinates in meters, which already falls within a physically possible scale. Therefore, scaling is not performed, but only parallel translation and rotation are performed.

Two types of preprocessing are performed in this study, as shown in Figure 3. The first type (i) is to reset the origin to the mean value of all observed trajectories $\mathcal{P}_{\text{obs}}$ by parallel transformation. This method keeps the actual distances between pedestrians and is used for analyzing causal relationships. The second (ii) is to rotate each trajectory and calculate the velocities at each time step. Previous studies have shown that predicting velocity yields more stable results than predicting locations. Based on this, to further alleviate the impact of the dominant direction in the training data, we propose rotating each trajectory so that the first movement points to the positive x direction, *i.e.*, making $x_n^1 \geq 0$ and $y_n^1 = 0$. It is used for node features and predictions for more stable results.

3.4 Causal-based Interaction Analysis

In this step, LiNGAM (Shimizu et al., 2006), a semi-parametric CD method, is employed to estimate the interactions among pedestrians. LiNGAM assumes a Linear function model, non-Gaussian noise distributions, and a directed Acyclic Graph. The non-Gaussian assumption ensures the identifiability. The following explains the causal model designed for this study and its limitations due to the assumptions.

The proposed causal model is shown in Eq. (1), and Figure 4 (left) illustrates an example of the resulting causal graph. The variables in LiNGAM should all be scalars, so each pedestrian is represented by two variables, (u_n, v_n). The causal model estimates a time-independent causal effect among pedestrians observed in the short observation period. Thus, the locations at each time step are a sample of observations, $\begin{bmatrix} u_1 & v_1 & \cdots & u_n & v_n \end{bmatrix}^T$. On the two sides of the equation, the symbol “:=” denotes the causal relationship, meaning only interfering with the right-hand side affects the left-hand side. $\mathbf{A}_{\text{Causal}}$ is the matrix of the parameters of linear functions to be estimated, which is also the edges of the causal graph. e_{u_n} and e_{v_n} are the noise (error) terms.

$$\begin{bmatrix} u_1 \\ v_1 \\ \vdots \\ u_n \\ v_n \end{bmatrix} := \mathbf{A}_{\text{Causal}} \begin{bmatrix} u_1 \\ v_1 \\ \vdots \\ u_n \\ v_n \end{bmatrix} + \begin{bmatrix} e_{u_1} \\ e_{v_1} \\ \vdots \\ e_{u_n} \\ e_{v_n} \end{bmatrix} \tag{1}$$

The linear assumption limits each pedestrian’s movement to a linear combination of the others who are the causes. The directed acyclic graph assumption means that no variable can be both a cause and an effect of another. It does not mean that pedestrians cannot be mutually affected, because the two variables representing the same person can have opposite causal effects. For example, the relationship between u_3, v_3, and u_2 in Figure 4 (left) conduct a bidirectional interaction between $n = 2$

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-4-2026-275-2026 | © Author(s) 2026. CC BY 4.0 License.

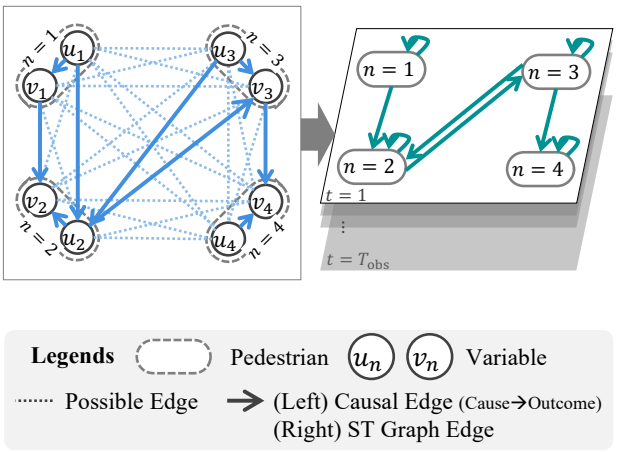


Figure 4. (Left) Example of a causal graph of four pedestrians Nine causal relationships are identified among all possible edges. (Right) Example of the ST graph based on the left. Node $n = 1$ has an edge to itself because of the causality $u_1 \rightarrow v_1$, and an edge pointing to $n = 2$ because $v_1 \rightarrow v_2$ and $u_1 \rightarrow u_2$.

and $n = 3$. Finally, to solve the $\mathbf{A}_{\text{Causal}}$, the number of observations (*i.e.*, the number of observation time steps) should be greater than the number of variables, which limits the number of pedestrians that can be estimated.

3.5 Building Spatial-Temporal Graph

The ST graph, $\mathcal{G}_{ST}$, is constructed as a directed graph to represent the causal relationship of observed pedestrians. Each node represents a pedestrian n at a given time step t and has two features of preprocessed coordinates (u_n^t, v_n^t). For each pedestrian, the edges are set between neighboring time steps. At each time step, the estimated causal graph determines the set of edges. Since two variables of a pedestrian are represented by separate nodes in the causal graph, the merging process is performed, resulting in an ST graph that is illustrated in Figure 4 (right). For any pair of pedestrians $n = j$ and $n = i$, if there exists a directed edge pointing from either u_j or v_j to either u_i or v_i in the causal graph, *i.e.*, j is a cause of i, then a directed edge pointing from node j to i is built in the ST graph. Because the CD step assumes causal relationships are independent of the short observation period, the edges between pairs of pedestrians are the same for all time steps.

3.6 Attention-Based Trajectory Prediction

Based on the ST graph, an attention-based DL model is developed to predict the future trajectories. The attention-based method learns explicit interactions between *each pair* of pedestrians. The proposed model adopts the GAT (Veličković et al., 2017) for learning interaction. The GAT is the simplest GNN that adopts an attention mechanism. Unlike GCNs, GATs can handle different weights for directional graphs. GATs also provide more interpretable weights because of the computation in the spatial domain, and the learned parameters are more flexible to graph structures. Our model follows the structure proposed in related studies (Mohamed et al. (2020); Qiu and Fuse (2024), *etc.*), composed of an interaction (spatial) feature extraction part, a temporal feature extraction part, and a temporal extrapolation part for generating predictions.

3.6.1 Interaction Feature Extraction: The ST graph is first passed to the interaction feature extraction module, which consists of several GATs. Eqs. (2)-(4) defines the computation of a GAT layer. In these equations, i, j, m index the nodes, $\vec{h_i}$ denotes the vector of features of node i, which is (u_i^t, v_i^t) in the first layer, and $\vec{h_i'}$ denotes the updated feature. Eq. (3) defines the “attention score” α_{ij}. It represents the degree to which node i pays attention to node j; in other words, how much i is affected by j. $\{E(j \rightarrow i) = 1\}$ denotes the set of nodes j that have a directed edge to node i, which means only causes of node i are accounted for their attention. The learnable parameters are $\boldsymbol{W}$, $\boldsymbol{W}_{\mathbf{GAT}}$, and $\vec{a}$. Through $\boldsymbol{W}$, the feature dimension is adjusted to consider variant features and adapt to the output. σ are nonlinear activation functions.

$$\vec{h_i'} = \sigma\left(\sum_{j\in\{E(j\rightarrow i)=1\}} \alpha_{ij}\boldsymbol{W}\vec{h_j}\right) \tag{2}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{m\in\{E(m\rightarrow i)=1\}}\exp(e_{im})} \tag{3}$$

$$e_{ij} = \sigma(\vec{a}^T[\boldsymbol{W}_{\mathbf{GAT}}\vec{h_i}||\boldsymbol{W}_{\mathbf{GAT}}\vec{h_j}]) \tag{4}$$

3.6.2 Temporal Feature Extraction & Extrapolation: Both of the steps are composed of CNN (Convolutional Neural Network) layers and nonlinear activation layers. Because the nodes are unordered, GNNs are not suitable for use, despite the temporal edges also being built. Otherwise, the time series of each pedestrian is seen as a sequence and learn by a pointwise CNN with a kernel size of three in the time dimension and one in the other dimensions.

3.6.3 Predicted Distribution & Loss function: Following the assumption by Bishop (1994) that the output of DL model is a Gaussian Mixture model, the output of our model are the five parameters of the two-dimensional Gaussian distribution of the future velocity in the 2D plane. In accordance with the assumption, the loss function is the probability of ground truth given the predicted probability distribution.

3.6.4 Sampling & Post-processing: Multiple possible trajectories are thus sampled from the output distribution and accumulated from velocity to trajectories.

4. Experiments& Discussion

4.1 Dataset and Implementation Details

The ETH (Pellegrini et al., 2009) and UCY (Lerner et al., 2007) datasets are used to evaluate this study. The dataset has served as a benchmark in numerous prior studies, enabling comprehensive comparisons with related work. The dataset primarily consists of pedestrians, with very few other agents or obstacles, such as bicycles or vehicles, making it suitable for studying interactions among pedestrians only. The dataset used for our experiments is obtained from Gupta (2018). The observation period T_{obs} is set as 23 frames (9.2s), and the prediction period T_{pred} is 12 frames (4.8s). The commonly used T_{obs} is 8 frames, so we re-implemented the methods of related studies to compare. The number of prediction samples, K, is 20, as is commonly used.

The models were trained with the stochastic gradient descent (SGD) optimizer, with a learning rate of 0.001 and a decay of 0.2 every 150 steps. The models with the lowest validation loss over 250 epochs were selected for testing.

4.2 Causal-Based Interaction Analysis

Figure 5(a) shows an example of nine pedestrians. The causal graph estimated by the CD step is shown in Figure 5(b), and Figure 5(c) shows the merged ST graph, corresponding to the left and right figures of Figure 4. Although there are no ground truths for causal relationships, interpreting and rationalizing the causal graph still gives us some insights into pedestrian behavior. It can be observed that interactions are not always present between all pairs of pedestrians. Most pedestrians are affected by nearby pedestrians but have no relationship with distant pedestrians. From this example, pedestrian $n = 6$ is neither a cause nor an outcome of $n = 4$ and 8, which is rational because the distances between them are always far. Additionally, $n = 6$ affects $n = 1$; this may be because $n = 1$ appears to proceed slowly to avoid conflict with $n = 6$, whereas $n = 6$ does not alter its approach because of $n = 1$. Although $n = 6$ was far from $n = 9$, they have a bidirectional effect, which might be because they are approaching each other. Additionally, some pedestrians, such as $n = 1$, usually have a one-way influence on the others; from the trajectory, $n = 1$ was almost static, and its location is likely to affect others. However, some causal relationships are difficult to explain, such as those between $n = 1$ and 4. Finally, even though the causal graph is restricted to a DAG, the ST graph still contains several bidirectional edges between pedestrians. Additional examples are shown in Figure 6 and discussed alongside the prediction.

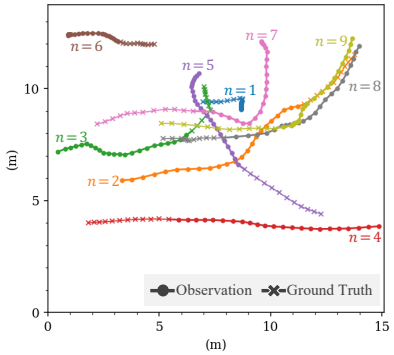
4.3 Attention-Based Trajectory Prediction

4.3.1 Accuracy Evaluation Metrics: minADE_{20} (Eq. (5)) and minFDE_{20} (Eq. (6)) defined by Alahi et al. (2016) are adopted for quantitatively evaluating the prediction results. Considering multiple possible future trajectories, these metrics account only for the minimum error among them. The average displacement error (ADE) is defined as the average error between the ground truth and predicted trajectories over all time steps, whereas the final displacement error (FDE) considers only the locations at the final time step. The symbol “$||\bullet||_2$” is defined as the Euclidean distance. K is the number of samples, usually set as 20, and indexed by k.

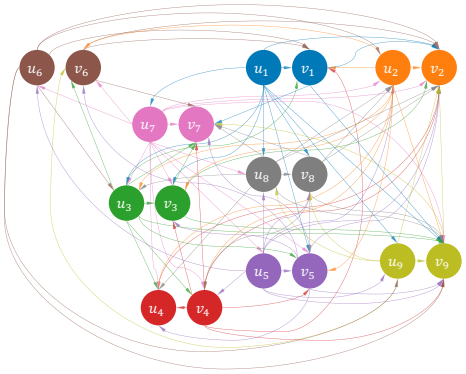
$$\text{minADE}_{\text{K}}(\hat{p_n^{t,k}}, p_n^{t,k}) = \min_{k\in\{1,\dots,K\}} \frac{\sum_{n=1}^{N}\sum_{t=T_{\text{obs}}+1}^{T_{\text{obs}}+T_{\text{pred}}} ||\hat{p_n^{t,k}} - p_n^{t,k}||_2}{N \times T_{\text{pred}}} \quad (5)$$

$$\text{minFDE}_{\text{K}}(\hat{p_n^{t,k}}, p_n^{t,k}) = \min_{k\in\{1,\dots,K\}} \frac{\sum_{n=1}^{N} ||\hat{p_n^{t,k}} - p_n^{t,k}||_2}{N}, \quad t = (T_{\text{obs}} + T_{\text{pred}}) \quad (6)$$

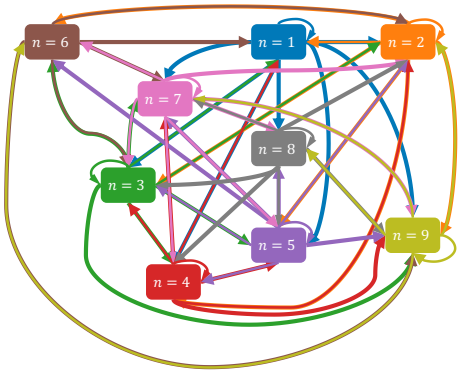
4.3.2 Quantitative Prediction Results: The quantitative results of predicted trajectories are shown in Table 1. The first row shows the proposed model. Two related studies in the second and third rows have similar model structures but are less interpretable. The model (a) used a complete graph, in which bidirectional edges are assumed to exist between all pairs of



(a) Ground truth trajectory: Observations are used for CD.



(b) Causal graph: Each node is colored with the same color in 5(a), and each edge is colored by the cause and point to the outcome.



(c) ST graph merged from 5(b)

Figure 5. Example containing nine pedestrians, from UNIV set

pedestrians. The model (b) adopted GCN and CNN for interaction modeling, which cannot distinguish interaction directions and may learn nonexistent relationships among neighbors from the training data. Although high-interpretability models usually result in lower accuracy, our prediction error is competitive with those of related studies. In the following two cases, our causal model is found to be ineffective when there are few pedestrians or many static pedestrians. First, our model performed worse than (b) on the ETH set, which has low pedestrian density. Under the DAG assumption and a few variables, the estimated causal edges must be sparse, even though the few pedestrians may have very close interactions. Second, our model performed the worst on the ZARA1 set, which contains many static pedestrians. When most variables representing movements have very

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-4-2026-275-2026 | © Author(s) 2026. CC BY 4.0 License.

Table 1. minADE_{20} and minFDE_{20} results

Model	minADE_{20} (unit: m)						minFDE_{20} (unit: m)					
	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
Ours	0.375	0.151	0.540	0.613	0.268	0.389	0.586	0.143	1.055	1.091	0.387	0.652
(a) Qiu and Fuse (2024)	0.488	0.161	0.520	0.455	0.257	0.376	0.741	0.165	0.957	0.820	0.359	0.608
(b) Mohamed et al. (2020)	0.260	0.317	0.701	0.547	0.497	0.465	0.214	0.348	1.092	0.833	0.522	0.602
(c) $\mathcal{E} = \{D < 1\,\mathrm{m}\}$	0.754	0.143	0.578	0.520	0.333	0.465	1.172	0.150	1.063	0.922	0.534	0.768
(d) $\mathcal{E} = \{D < 2\,\mathrm{m}\}$	0.764	0.139	0.561	0.449	0.281	0.439	1.240	0.144	1.034	0.777	0.440	0.727
(e) $\mathcal{E} = \{D < 5\,\mathrm{m}\}$	0.382	0.138	0.638	0.448	0.252	0.371	0.522	0.135	1.131	0.764	0.374	0.585

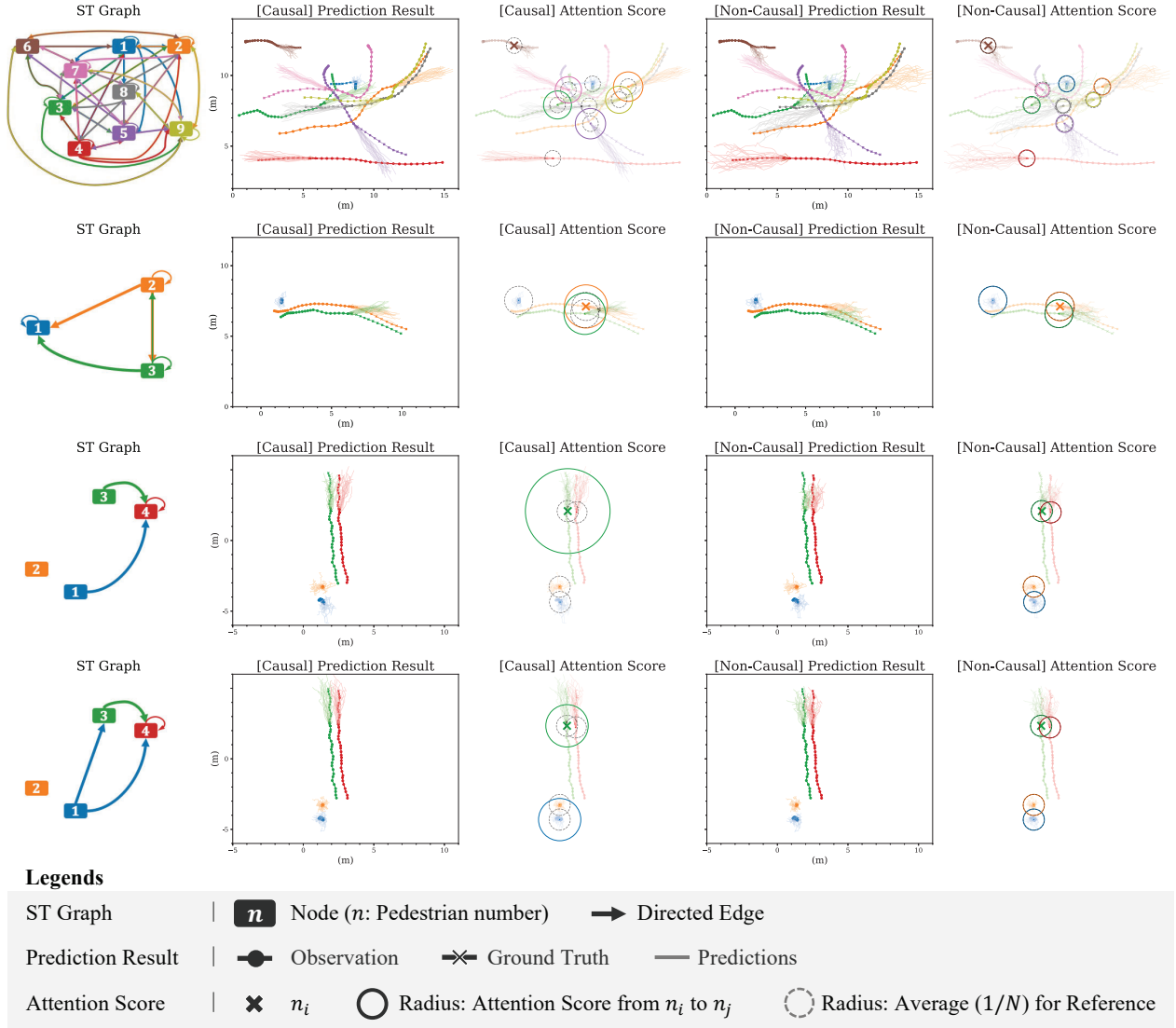


Figure 6. Results of CD and Trajectory Prediction: The first column visualizes the estimated and merged ST graphs; the second and third columns are the prediction results and a subset of examples of the attention scores from our proposed CD-based approach; the last two columns are those from a non-causal ablation study. The ST graph of the first row is same as Figure 5(c)

small magnitudes, the causal discovery method may estimate unstable edges. Namely, past movements provide insufficient clues about the causal relationships.

4.3.3 Ablation study: The model (a) can be considered as the non-causal version of our proposed approach because all the edges exist. The models (c)–(e) consider the existence of the edges depending on the current physical distance rather than past movements; only pedestrians within a threshold of specific distances are assumed to interact with each other. The accuracy of (c)–(e) is seldom better than ours or (a). Moreover, it is challenging to determine a threshold across different scenes, and the distance threshold cannot account for unidirectional interaction. Thus, this comparison supports that determining edges based on causality is better than using a fixed threshold.

4.4 Visualization & Discussion

In Figure 6, the first column shows the estimated and merged ST graphs, and the following two pairs of columns are the res-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-4-2026-275-2026 | © Author(s) 2026. CC BY 4.0 License.

ults and attention scores of our proposed method ([Causal]) and the model (a)([Non-Causal]). Each row shows a piece of data, which can also be considered as a graph $\mathcal{G}_{ST}$.

Regarding the prediction results, the multiple predictions for the causal ones were more concentrated than those for the non-causal ones. This implied that the causal results were more confident, with smaller variances. However, the second row shows that making predictions in the wrong direction can lead to larger errors in minADE_{20} and minFDE_{20}. Conversely, these error metrics have the limitation that they can be small when the variances are large, implying large uncertainty. The finding is also supported by other studies (Ivanovic and Pavone, 2019; Mohamed et al., 2022).

Regarding the effect of the ST graphs on the results, the first row shows a promising outcome, indicating that edge pruning can lead to better results, especially in complex situations. Comparing the attention scores of the causal ones to the non-causal ones reveals that learning GAT with a complete graph tends to result in the same scores for all nodes, which is related to the over-smoothing problem of many GNN methods (Chen et al., 2020). Although designing more complex models, such as using Graphormer (Ying et al., 2021), can overcome this problem, they again sacrifice interpretability.

Unfavorable results of the causal-based ST graph are also identified. The second row shows that cutting too many edges would result in insufficient attention. In this case, the orange pedestrian ($n = 2$) is probably affected by the other two pedestrians, whereas the causal-based attention is only on itself and the green one ($n = 3$). Due to the small number of pedestrians and the DAG assumption, the weakness of edge insufficiency is identified from both the quantitative and the visualization results. Nonetheless, the DAG assumption is difficult to relax because the identifiability of causal models relies on the DAG. If the edges can form a loop, it is hard to tell which is the actual cause. Also, the example of the last two rows presents the instability of CD. Although the situations in the last two rows are very similar, different causal graphs are estimated and result in very different attention scores of the green ($n = 3$) pedestrian. It reflects the difficulty of balancing the need for more observations to achieve stable CD results with the desire to shorten the observation period to provide more real-time suggestions. Besides, a small amount of movement also causes instability in the CD results, as discussed in the quantitative results section.

Some limitations not shown in the figures are also identified. First, excluding the environmental factors enables this study to focus on pedestrian interactions, but some actions of people heading to a garbage can or avoiding a tree cannot be predicted. Our related work (Qiu and Hato, 2025) discusses modeling environmental factors by considering object-level interactions between them and pedestrians. Second, the combination of the linear causal relationship from LiNGAM and the static CD model design cannot capture changes in causal relationships over time. Although most relationships between pedestrians remain unchanged within the short observation window (9.2 seconds in our experiments), some changes in relationships are observed after pedestrians pass each other and disappear from each other's field of view. This may be improved by adopting nonlinear CD methods and designing a dynamic CD model, but the number of observations should increase accordingly.

5. Conclusions

In this study, we developed a pedestrian behavior modeling method that strikes a good balance between relatively high prediction accuracy and interpretability by integrating both CD and attention-based DL methods. The CD step allowed causal interpretations, which are stronger interpretations than correlations. Also, visualizing the causal-based ST graphs and the explicitly learned attention scores improved our understanding of how the model manipulates interactions. From the experiments, we explored the characteristics of CD and prediction results and observed potential issues. Future work includes finding a better balance between the number of pedestrians and the number of observations to provide stable CD results and sufficient edges, considering nonlinear and dynamic causal relationships among pedestrians, and considering environmental factors. Finally, the quantitatively estimated interactions are expected to benefit the simulation of pedestrians and autonomous agents.

References

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S., 2016. Social LSTM: Human trajectory prediction in crowded spaces. *2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 961–971.

Antonini, G., Bierlaire, M., Weber, M., 2006. Discrete choice models of pedestrian walking behavior. *Transp. Res. Part B: Methodol.*, 40(8), 667–687.

Bishop, C. M., 1994. Mixture density networks. Technical Report, https://publications.aston.ac.uk/id/eprint/373/.

Blue, V., Adler, J., 1998. Emergent fundamental pedestrian flows from cellular automata microsimulation. *Transportation research record*, 1644, 29–36.

Brouwer, N., Kloeden, H., Stiller, C., 2016. Comparison and evaluation of pedestrian motion models for vehicle safety systems. *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, 2207–2212.

Chen, D., Lin, Y., Li, W., Li, P., Zhou, J., Sun, X., 2020. Measuring and Relieving the Over-Smoothing Problem for Graph Neural Networks from the Topological View. *Proc. of the AAAI Conference on Artificial Intelligence*, 34(04), 3438-3445.

Daamen, W., 2004. Modelling Passenger Flows in Public Transport Facilities. PhD thesis, Delft University of Technology.

Gupta, A., 2018. Social GAN. Github Code, github.com/agrimgupta92/sgan.

Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A., 2018. Social GAN: Socially acceptable trajectories with generative adversarial networks. *2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2255–2264.

Helbing, D., Farkas, I. J., Molnar, P., Vicsek, T., 2002. Simulation of pedestrian crowds in normal and evacuation situations. *Pedestrian and Evacuation Dynamics*, 21–58.

Hoogendoorn, S. P., Bovy, P. H., 2002. Normative pedestrian behaviour theory and modelling. *Transportation and Traffic Theory in the 21st Century*, Emerald Group Publishing Limited, Adelaide, Australia, 219–245.

Huang, Y., Bi, H., Li, Z., Mao, T., Wang, Z., 2019. STGAT: Modeling spatial-temporal interactions for human trajectory prediction. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6271–6280.

Ivanovic, B., Pavone, M., 2019. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. *Proceedings of the IEEE/CVF international conference on computer vision*, 2375–2384.

Jain, A., Zamir, A. R., Savarese, S., Saxena, A., 2016. Structural-RNN: Deep learning on spatio-temporal graphs. *2016 IEEE Conf. Comput. Vis. Pattern Recognit.*, 5308–5317.

Karamouzas, I., Heil, P., van Beek, P., Overmars, M. H., 2009. A predictive collision avoidance model for pedestrian simulation. *Motion in Games*, 41–52.

Kosaraju, V., Sadeghian, A., Martín-Martín, R., Reid, I., Rezatofighi, H., Savarese, S., 2019. Social-BiGAT: Multimodal trajectory forecasting using Bicycle-GAN and graph attention networks. *Adv. Neural Inf. Process.*, 32.

Kothari, P., Sifringer, B., Alahi, A., 2021. Interpretable social anchors for human trajectory forecasting in crowds. *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 15556–15566.

Lerner, A., Chrysanthou, Y., Lischinski, D., 2007. Crowds by Example. *Computer Graphics Forum*, 26(3), 655–664.

Makansi, O., von Kügelgen, J., Locatello, F., Gehler, P. V., Janzing, D., Brox, T., Schölkopf, B., 2021. You mostly walk alone: Analyzing feature attribution in trajectory prediction. arxiv.org/abs/2110.05304.

Malinsky, D., Spirtes, P., 2016. Estimating causal effects with ancestral graph Markov models. *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*, Proceedings of Machine Learning Research, 52, 299–309.

Minoura, H., Hirakawa, T., Yamashita, T., Fujiyoshi, H., 2022. Trajectory Forecasting Considering Interactions between Moving Targets Using DeepLearning: A Survey. *IEICE TRANSACTIONS on Information and Systems*, J105-D(5), 372–404.

Mohamed, A., Qian, K., Elhoseiny, M., Claudel, C., 2020. Social-STGCNN: A social spatio-temporal graph convolutional neural network for human trajectory prediction. *2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 14412–14420.

Mohamed, A., Zhu, D., Vu, W., Elhoseiny, M., Claudel, C., 2022. Social-Implicit: Rethinking trajectory prediction evaluation and the effectiveness of implicit maximum likelihood estimation. *Computer Vision – ECCV 2022*, 463–479.

Papadimitriou, E., Yannis, G., Golias, J., 2009. A critical assessment of pedestrian behaviour models. *Transp. Res. Part F: Psychol. Behav.*, 12(3), 242-255.

Pellegrini, S., Ess, A., Schindler, K., van Gool, L., 2009. You’ll never walk alone: Modeling social behavior for multi-target tracking. *2009 IEEE 12th International Conference on Computer Vision*, 261–268.

Qiu, W.-X., Fuse, T., 2024. Visualization of Pedestrian Interaction through Attention-based Pedestrian Trajectory Prediction. *Asain Journal of Geoinformatics*, AJG-2311002.

Qiu, W.-X., Hato, E., 2025. Pedestrian-object interaction modeling through heterogeneous graph attention networks. *72nd Conference, Autumn Conference of Infrastructure Planning and Management, JSCE*.

Robin, T., Antonini, G., Bierlaire, M., Cruz, J., 2009. Specification, estimation and validation of a pedestrian walking behavior model. *Transp. Res. Part B: Methodol.*, 43(1), 36–56.

Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Rezatofighi, H., Savarese, S., 2019. SoPhie: An attentive GAN for predicting paths compliant to social and physical constraints. *2019 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 1349–1358.

Sani, N., Malinsky, D., Shpitser, I., 2023. Explaining the behavior of black-box prediction algorithms with causal learning. arxiv.org/abs/2006.02482.

Schadschneider, A., 2002. Cellular automaton approach to pedestrian dynamics-theory. *Pedestrian and Evacuation Dynamics*.

Shafiee, N., Padir, T., Elhamifar, E., 2021. Introvert: Human trajectory prediction via conditional 3D attention. *Proc. of the IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 16815–16825.

Shi, L., Wang, L., Long, C., Zhou, S., Zhou, M., Niu, Z., Hua, G., 2021. SGCN: Sparse graph convolution network for pedestrian trajectory prediction. *2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 8990–8999.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., 2006. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research*, 7, 2003–2030.

Spirtes, P., Glymour, C. N., Scheines, R., 2000. *Causation, prediction, and search*. 2 edn, MIT press.

Sun, H., Zhao, Z., He, Z., 2020. Reciprocal learning networks for human trajectory prediction. *2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, IEEE, 7414–7423.

Unhelkar, V. V., Pérez-D’Arpino, C., Stirling, L., Shah, J. A., 2015. Human-robot co-navigation using anticipatory indicators of human walking motion. *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 6183–6190.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. *Adv. Neural Inf. Process. Syst.*, 30, 6000–6010.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y., 2017. Graph attention networks. arxiv.org/abs/1710.10903.

Vemula, A., Muelling, K., Oh, J., 2018. Social Attention: Modeling attention in human crowds. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 4601–4607.

Xu, Y., Piao, Z., Gao, S., 2018. Encoding crowd interaction with deep neural network for pedestrian trajectory prediction. *2018 CVPR*, 5275–5284.

Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., Liu, T.-Y., 2021. Do transformers really perform badly for graph representation? *Adv. Neural Inf. Process. Syst.*, 34, 28877–28888.

Yu, C., Ma, X., Ren, J., Zhao, H., Yi, S., 2020. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. *Computer Vision – ECCV 2020*, 507–523.