Causal Intervention for Human Trajectory Prediction with Cross Attention Mechanism

Chunjiang Ge, Shiji Song, Gao Huang*

Department of Automation, BNRist, Tsinghua University
gecj20@mails.tsinghua.edu.cn, {shijis, gaohuang}@tsinghua.edu.cn

Abstract

Human trajectory Prediction (HTP) in complex social environments plays a crucial and fundamental role in artificial intelligence systems. Conventional methods make use of both history behaviors and social interactions to forecast future trajectories. However, we demonstrate that the social environment is a confounder that misleads the model to learn spurious correlations between history and future trajectories. To end this, we first formulate the social environment, history and future trajectory variables into a structural causal model to analyze the causalities among them. Based on causal intervention rather than conventional likelihood, we propose a **S**ocial **E**nvironment **AD**justment (SEAD) method, to remove the confounding effect of the social environment. The core of our method is implemented by a **S**ocial **C**ross **A**ttention (SCA) module, which is universal, simple and effective. Our method has consistent improvements on ETH-UCY datasets with four baseline methods and achieves competitive performances with existing methods.

Introduction

Human trajectory prediction is a fundamental and essential task for several social applications, such as intelligent transport systems and socially-aware robotic navigation (Chandra et al. 2019; Liang et al. 2019). For example, autonomous driving systems rely on accurate future trajectory prediction to control the vehicles and avoid collisions (Bai et al. 2015; Morotomi, Katoh, and Hayashi 2014). In addition, the human trajectory prediction models could be used by surveillance systems to identify pedestrians(Luber et al. 2010; Musleh et al. 2010). Due to these important applications, human trajectory prediction methods have been extensively investigated in the literature.

The prevailing pipeline to train trajectory prediction model is shown in Figure 1 (a). History trajectory $\mathbf{X}$ and the surrounding pedestrians $\tilde{\mathbf{X}}$ are incorporated to predict future trajectory $\mathbf{Y}$. Existing methods focus on modeling the social interactions with history behaviors for prediction (Kosaraju et al. 2019; Pellegrini et al. 2009). Social-LSTM leverages a social pooling module to exploit social environments (Alahi et al. 2016). STGAT (Huang et al. 2019) and Social-STGCNN (Mohamed et al. 2020) apply graph neural networks to aggregate social information by introducing graph attention and weighted adjacency matrix.

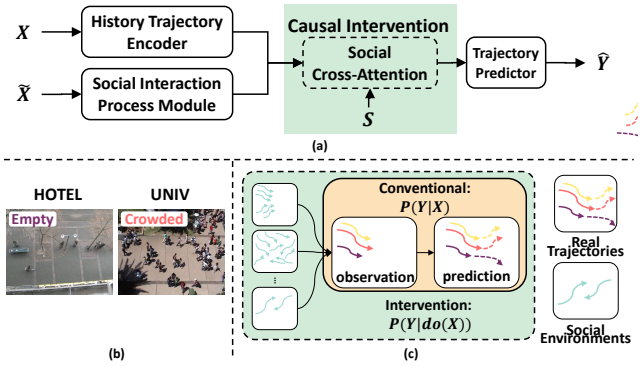

Figure 1: (a) The general pipeline of the trajectory prediction models (beyond the green block) (Huang et al. 2019; Mohamed et al. 2020; Liu et al. 2022). We propose to use a Social Self-Attention module to leverage causal intervention to adjust the social environment (in the green block). (b) The comparison between *HOTEL* and *UNIV*. We could observe the social distances between pedestrians are distinct between these two domains. (c) The illustration of how our SEAD method works. With causal intervention, $P(\mathbf{Y}|do(\mathbf{X}))$ incorporates every social environment pattern into the scene to predict future trajectories. Compared with our method, conventional methods are shown in the yellow box which predict future trajectories with likelihood predictor $P(\mathbf{Y}|\mathbf{X})$.

In spite of the progress, human trajectory prediction still remains a challenging problem since the social environments are complex and biased (Liu, Yan, and Alahi 2021; Chen et al. 2021). The trajectory data are collected at various places and times whose social environments are distinct (Amirian et al. 2020; Lerner, Chrysanthou, and Lischinski 2007; Ge et al. 2022). The social motion patterns, social interactions *etc.* are different (Amirian et al. 2020; Pellegrini et al. 2009): *e.g.*, people tend to form groups in the crowded UNIV domain since the scene is crowded, while pedestrians mostly keep a social distance from each other in HOTEL domain(Pellegrini, Ess, and Gool 2010) (Figure 1 (b)). It's worth noting that the biased social environment features lead to spurious correla-

*Corresponding Author.

tions. For example, people may proceed in parallel with the same speed since they are in a crowded environment rather than forming a group. The prediction could be inaccurate since the model wrongly correlates a crowded environment to the group motion pattern.

Based on the above analysis, we have observed how the social environment influences trajectory prediction. Essentially, the social environment is a confounder that misleads the model to learn the spurious correlations between history and predicted trajectories. The confounding effect could not be eliminated by current trajectory prediction methods without causal intervention (Pearl 2009). To systematically investigate the problem, we formulate the causalities among social environments, history and future trajectories into a Structural Causal Model (SCM) (Pearl 2009).

Based on such SCM, we propose a causal intervention method, named Social Environment ADjustment (SEAD) to remove the confounding effect of social environments through backdoor adjustment (Pearl 2009). The key difference from conventional methods is that SEAD aims at learning the interventional probability rather than conventional likelihood. However, we could not directly apply causal intervention on history trajectory $\mathbf{X}$, since the social environment $\mathbf{S}$ is not well-defined. Hence, we propose to design a dictionary $\{\mathbf{s}_i\}$ to approximate the representation of the social environment $\mathbf{S}$ and perform backdoor adjustment on it.

To implement the key of our idea, we propose a simple yet effective Social Cross Attention (SCA) module to realize the causal intervention on trajectory features. The SCA module leverages external social environment variables with internal trajectory features. The learnable social environment variables are encoded into keys and values to interact with queries of trajectory features. It could be intuitively understood as follows (Figure 1 (c)): the module incorporates every social environment pattern into the scene to predict future trajectories. As a consequence, the spurious correlations between history and future trajectories are cut off.

We demonstrate the effectiveness of SEAD on trajectory prediction dataset ETH (Pellegrini, Ess, and Gool 2010) and UCY (Lerner, Chrysanthou, and Lischinski 2007). Our SEAD method could be applied to both RNN-based and CNN-based frameworks, including STGAT (Huang et al. 2019), Social-STGCNN (Mohamed et al. 2020) and TF (Giuliari et al. 2021). Additionally, our method could further improve the performance of Causal-STGAT (Chen et al. 2021) since they are orthogonal. We show that our method achieves consistent improvements on both four baseline models. With SEAD, STAGT, Social-STGCNN, TF and Causal-STGAT have an improvement of $0.05/0.09$, $0.03/0.04$, $0.04/0.08$ and $0.03/0.07$ on the ADE/FDE metrics respectively. Through qualitative experiments, we demonstrate that our method is able to produce more reasonable trajectories for pedestrians.

Related Work

Trajectory Prediction. Temporal information and social information are two critical factors in predicting feature trajectories (Huang et al. 2021). For temporal information, the future sequence could be predicted through learning the underlying temporal associations from the history sequence. Recurrent Neural Networks (RNN), especially Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber 1997), are designed for modeling such information. Besides, Nikhil *et al.* proposed to use temporal Convolutional Neural Networks (CNN) to predict future trajectories(Nikhil and Morris 2018).

For social information, traditional methods leverage physical and pre-defined rules to exploit social interactions, *e.g.*, Social Force Model (Helbing and Molnar 1995) and Discrete Choice framework (Antonini, Bierlaire, and Weber 2006). In deep learning methods, social pooling is proposed for social interaction aggregation with equal weights(Alahi et al. 2016). A line of methods applies attention mechanism to distinguish the importance of different neighboring pedestrians (Kosaraju et al. 2019; Sadeghian et al. 2019). Graph Neural Networks (GNN) are used for trajectory prediction since they could aggregate the social information according to the pre-defined metrics, *e.g.*, distance between pedestrians (Huang et al. 2019; Mohamed et al. 2020; Kosaraju et al. 2019). These conventional methods are based on likelihood $P(Y|X)$ while we propose to use causal intervention to model the causalities $P(Y|do(X))$.

Attention Mechanism. Attention mechanism is first proposed in natural language processing to better model long-term sequential relations (Vaswani et al. 2017; Pan et al. 2022). Features are mapped into queries, keys and values to model the interactions between tokens. Attention mechanism is used in computer vision and multi-modality tasks recently (Dosovitskiy et al. 2021; Radford et al. 2021; Li et al. 2021). Cross attention module(Vaswani et al. 2017) is widely used to fuse different feature maps (Hao et al. 2017; Gheini, Ren, and May 2021). In our proposed method, the input of keys and values are trainable social environment variables to perform causal intervention.

Causal Inference. Causal inference (Imbens and Rubin 2015; Pearl 2009) is developed to estimate causal effect with covariate shift (Richiardi, Bellocco, and Zugna 2013; Glymour, Pearl, and Jewell 2016). By causal intervention, the spurious correlations between cause and effect are cut off and the true causality could be accurately estimated (Glymour, Pearl, and Jewell 2016). Many methods taking inspiration from causal inference have been explored to help deep neural networks to learn the true causalities (Zhang et al. 2020; Tang, Huang, and Zhang 2020). The true causalities could reduce the prediction error of deep learning models (Niu et al. 2021). In this paper, we propose to use causal intervention to remove the spurious correlations.

A related work to us is *Counterfactual Analysis* (Chen et al. 2021). It uses counterfactual analysis to reduce the discrepancy between training and deployment environments. The main differences are as follows. First, *Counterfactual Analysis* applies mediation analysis which is still based on likelihood, while we leverage causal intervention method (Richiardi, Bellocco, and Zugna 2013). Second, *Counterfactual Analysis* aims to alleviate the negative effect of the discrepancy between training and deployment environments, while our method resorts to dealing with the biased social environment. Last, *Counterfactual Analysis* maintains the relations between pedestrians, while our method adjusts

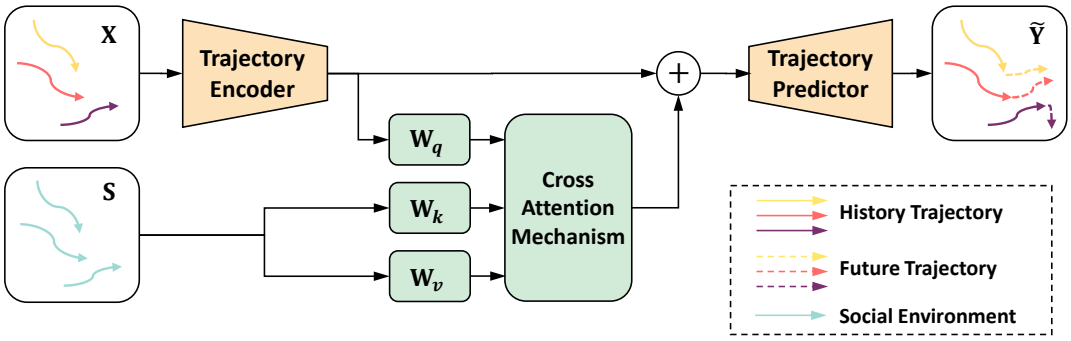

Figure 2: An overview of our proposed Social Environment Adjustment (SEAD) method. The color yellow stands for conventional models and blue stands for our method. $\mathbf{W}_q, \mathbf{W}_k.\mathbf{W}_v$ are linear transformation blocks.

the biased relations via causal intervention. In addition, another work (Liu et al. 2022) considers the robustness of HTP models with respect to noisy input while we focus on the confounding bias.

Causal Attention for Social Interaction

The trajectory prediction task is defined as a sequential prediction problem with history trajectories and social interactions as input. Given m pedestrians in a scene, the history trajectory of the i-th pedestrian is defined as $\mathbf{x}_i = \{\mathbf{p}_i^t = (x_i^t, y_i^t) \in \mathcal{R}^2 | t = 1, 2, \ldots, T_{\mathrm{h}}\}$, where the (x_i^t, y_i^t) is the 2D location of the pedestrian at time t and the T_{h} is the length of the history trajectory. The ground truth future trajectory is $\mathbf{y}_i = \{\mathbf{p}_i^t = (x_i^t, y_i^t) \in \mathcal{R}^2 | t = T_{\mathrm{h}}+1, T_{\mathrm{h}}+2, \ldots, T_{\mathrm{h}}+T_{\mathrm{f}}\}$, where T_{f} is the length of ground truth future trajectory. A line of methods employs the relative locations or velocity for future trajectory prediction (Huang et al. 2019; Mohamed et al. 2020). The social interaction information for i-th pedestrian could be defined as a function of surrounding pedestrians' trajectories $\mathbf{e}_i = e(\tilde{\mathbf{x}}_i)$, where $e(\cdot)$ is a function for aggregating social information (*e.g.*, social pooling in Social-LSTM) and $\tilde{\mathbf{x}}_i = [\mathbf{x}_j]_{j=0}^{m, j\neq i}$. The formalization for HTP is:

$$\hat{\mathbf{y}}_i = f(\mathbf{x}_i, \mathbf{e}_i), \qquad L = \sum_{i=0}^{N} \mathcal{L}(\mathbf{y}_i, \hat{\mathbf{y}}_i), \tag{1}$$

where the $\hat{\mathbf{y}}_i$ is the future trajectory predicted by the model $f(\cdot)$ and $\mathcal{L}(\cdot)$ is the loss function.

As discussed in Introduction, the social environment $\mathbf{s}$ has a confounding effect on predicting the future trajectories $\mathbf{y}$. Next, we would introduce a causal graph to elaborate on how the biased social environment would impact trajectory prediction and how to estimate the true causal effect. At last, we investigate into the essence of backdoor adjustment and devise a Social Cross-Attention module to make precise future trajectory prediction.

Structural Causal Model

We construct the structural causal model for trajectory prediction to analyze the causal relations. As shown in Figure 3 (a), the structural causal model contains three variables (nodes): $\mathbf{X}$: history trajectory, $\mathbf{Y}$: future trajectory and $\mathbf{S}$: social environment. The directed edges in the structural causal model

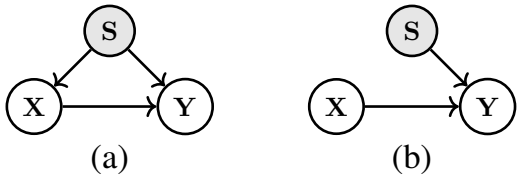

Figure 3: The proposed structural causal model of our method. The gray color denotes $\mathbf{S}$ is unobserved. (a) The original SCM. (b) The SCM after the causal intervention.

represent causalities between two nodes: cause $\rightarrow$ effect. We would discuss these directed edges in the graph.

$\mathbf{X} \rightarrow \mathbf{Y}$. The future trajectory could be inferred by the clues of history trajectory. The history trajectory contains the following information: velocity, acceleration and potential beginning position of the future trajectory. For example, pedestrians would maintain proceed in a line rather than changing the direction frequently.

$\mathbf{S} \rightarrow \mathbf{Y}$. The social environment would influence the future trajectory. The social environment influences the mode of pedestrians. For example, pedestrians would change their direction when other pedestrians walk to avoid collisions. Besides, people tend to walk parallel to maintain the distances between them (Lerner, Chrysanthou, and Lischinski 2007).

$\mathbf{S} \rightarrow \mathbf{X}$. The history trajectory is affected by the social environment for the same reason with $\mathbf{S} \rightarrow \mathbf{Y}$. We need to point out that the social environment could vary with time goes. For example, pedestrians may walk alone in the past and in parallel with other pedestrians in the future.

Based on the above analysis, we could point out that the social environment plays the role of confounder in the trajectory prediction task (Imbens and Rubin 2015). We could see clearly that there exists a backdoor path $\mathbf{X} \leftarrow \mathbf{S} \rightarrow \mathbf{Y}$ in the Figure 3 (a). The backdoor path indicates that even if some history trajectories $\mathbf{X}$ have little likelihood to produce some unreasonable future trajectories, the social environment could still correlate $\mathbf{X}$ with $\mathbf{Y}$, leading to the problem mentioned in Introduction. Learning $P(\mathbf{Y}|\mathbf{X})$ from the biased social environments would include the spurious correlation in the model. Therefore, we propose a causal intervention method to remove the confounding effect. With intervention on $\mathbf{X}$, the directed edge $\mathbf{S} \rightarrow \mathbf{X}$ is cut off and spurious correlations are eliminated (Figure 3).

Causal Intervention via Backdoor Adjustment

We propose to use causal intervention $P(\mathbf{Y}|do(\mathbf{X}))$, named Social Environment ADjustment (SEAD), to remove the confounding effect of social environment $\mathbf{S}$. The core idea of our SEAD method is (a) applying backdoor adjustment to remove the confounding effect, and (b) leveraging a set of learnable variables $\{\mathbf{s}_i\}_{i=1}^n$ to approximate the social environment.

To learn the probability $P(\mathbf{Y}|do(\mathbf{X}))$, we incorporate backdoor adjustment to perform intervention on the input history trajectory $\mathbf{X}$ (Figure 3 (a)). However, backdoor adjustment requires social environment $\mathbf{S}$ could be stratified. The social environment does not have a strict and well-accepted definition, let alone stratifying it. To end this, we propose to use a set of learnable variables $\{\mathbf{s}_i\}$ to approximate the social environment $\mathbf{S}$ automatically. Such approximation strategy implies we discretize the representation of the social environment and assume there are several "typical" patterns of the social environment. This assumption is reasonable because the social environment is subject to several factors, *e.g.*, social norms, physical rules and certain scenes (Liu et al. 2022; Helbing and Molnar 1995). Here, we could introduce the backdoor adjustment (the detailed derivation is in Appendix):

$$P(\mathbf{Y}|do(\mathbf{X})) = \sum_{i}^{n} P(\mathbf{Y}|\mathbf{X}, \mathbf{S} = \mathbf{s}_i)P(\mathbf{s}_i), \tag{2}$$

where the n is the length of the set of learnable social environment $\{\mathbf{s}_i\}_{i=1}^n$ and each $\mathbf{s}_i$ is a d-dimensional vector. $P(\mathbf{Y}|\mathbf{X})$ is essentially different from $P(\mathbf{Y}|do(\mathbf{X}))$. According to the law of total probability, $P(\mathbf{Y}|\mathbf{X})$ has the conditional probability $P(\mathbf{s}_i|\mathbf{X})$ rather than the marginal probability $P(\mathbf{s}_i)$. According to Eq. (2), the future trajectory is calculated by incorporating every social environment $\mathbf{s}_i$ into the scene with weights $P(\mathbf{s}_i)$. Hence, each pattern of the social environment would be considered (subject to the prior $P(\mathbf{s}_i)$) in the prediction and the spurious correlation would not dominate. Specifically, the backdoor path $\mathbf{X} \leftarrow \mathbf{S} \rightarrow \mathbf{Y}$ is cut off and social environment $\mathbf{S}$ and history trajectory $\mathbf{X}$ would be independent (Figure 3 (b)).

Eq. (2) requires the estimation of $P(\mathbf{Y}|\mathbf{X}, \mathbf{S} = \mathbf{s}_i)$ and $P(\mathbf{s}_i)$. However, there exists difficulties in strictly estimating the high dimensional conditional distribution $P(\mathbf{Y}|\mathbf{X}, \mathbf{S})$ and a probability distribution dependent on the learnable variables $\{\mathbf{s}_i\}$. Considering HTP requires 2D trajectory prediction, we propose to predict future trajectories with the intervention: $\hat{\mathbf{y}} = \sum_{i=1}^{n} f_y(\mathbf{x}, \mathbf{s}_i)P(\mathbf{s}_i)$, compared with conventional methods: $\hat{\mathbf{y}} = f(\mathbf{x})$. We would formulize the procedure of predicting the future trajectories $\hat{\mathbf{y}}$ with intervention into a Social Cross Attention module.

Social Cross-Attention Module

Since $f_y(\mathbf{x}, \mathbf{s}_i)$ requires modeling the features of both trajectory and social environment, we propose a Social Cross Attention (SCA) module to fuse these features because cross attention is adopted at modeling the connections between trajectory $\mathbf{x}$ and social environment $\mathbf{s}$. Different from self-attention, our Social Cross Attention module incorporates both internal trajectory variables $\mathbf{x}$ and external model variables $\mathbf{s}$. Hence, the overall implementation is:

$$\mathrm{SCA}(\mathbf{x}, \mathbf{s}) = \mathbb{E}_{\mathbf{s}}[(f_y(\mathbf{x}, \mathbf{s}))]. \tag{3}$$

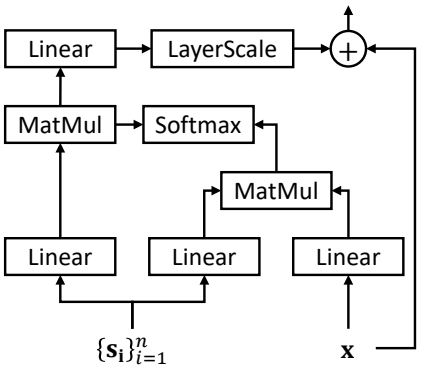

Figure 4: The structure of our proposed Social Cross Attention module.

We tailor backdoor adjustment into HTP by introducing the details. The trajectory features and social environment variables are first projected into the same space. Then the similarity between $\mathbf{x}$ and $\mathbf{s}$ is calculated by matrix multiplication. Finally, we aggregate the features with the similarity matrix and project the feature maps into the original space. Specifically, the Social Cross Attention module is formulated as follows (Figure 4):

$$\mathbf{q} = \mathbf{W}_q\mathbf{x}, \mathbf{k}_i, \mathbf{v}_i = \mathbf{W}_k\mathbf{s}_i, \mathbf{W}_v\mathbf{s}_i, \tag{4}$$

$$\mathbf{z} = \mathbb{E}_{\mathbf{s}}[f_y(\mathbf{x}, \mathbf{s})] = \mathbf{W}_o \sum_{i}^{n} \mathrm{Softmax}(\frac{\mathbf{q}^T\mathbf{k}_i}{\sqrt{d}})\mathbf{v}_i P(\mathbf{s}_i), \tag{5}$$

where $\mathbf{z}$ denotes the output tensor of SCA and $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$ are learnable linear transformation of SCA to project $\mathbf{x}$ and $\mathbf{s}$ into the same space. The linear transformation $\mathbf{W}_o$ is applied to project the output into the 2D location space. The scaling factor is $\sqrt{d}$ here. The LayerScale is applied for convergence and we omit it for brevity in Eq. (5).

According to Eq. (5), our SCA module requires the distribution of $P(\mathbf{s})$. However, we do not have knowledge of the social environment distribution of each group $P(\mathbf{s}_i)$. Hence, we assume that each group $\mathbf{s}_i$ could be *fairly* incorporated into each scene. The probability of social environment follows a uniform distribution $P(\mathbf{s}_i) = 1/n$. Besides, if $P(\mathbf{s})$ does not subject to the dataset statistics, our method could generalize well.

According to the physical information in the feature maps, our SCA module incorporates multi-head attention mechanism. The feature maps for STGAT is designed for forecasting 2D locations and Social-STGCNN for 2D locations, variance and covariance. Specifically, the feature map is divided into 2 heads for STGAT and 5 heads for Social-STGCNN. Finally, the output of SCA $\mathbf{z}$ is fused with trajectory features $\mathbf{x}$ and the overall prediction for the future trajectory is:

$$\hat{\mathbf{y}} = h(\mathbf{x} + \mathbb{E}_{\mathbf{s}}[f_y(\mathbf{x}, \mathbf{s}))] = h(\mathbf{x} + \mathbf{z}), \tag{6}$$

where $h(\cdot)$ represents the trajectory predictor.

Experiments

We evaluate our proposed SEAD in terms of performance quantitatively. The datasets, evaluation metrics and implementation details are introduced. We demonstrate the effec-

Setting		Performance (ADE/FDE)					
		ETH	HOTEL	ZARA1	ZARA2	UNIV	AVG
(**Q1**)	STGAT	0.83/1.74	0.31/0.59	0.39/0.85	0.31/0.68	0.50/1.08	0.47/0.99
	MHSA	0.82/1.74	0.32/0.61	0.39/0.86	0.32/0.70	0.50/1.08	0.47/1.00
	SCA	0.68/1.40	0.29/0.56	0.35/0.79	0.30/0.70	0.48/1.04	0.42/0.90
	Social-STGCNN	0.67/1.13	0.40/0.62	0.34/0.53	0.30/0.48	0.52/0.96	0.45/0.75
	MHSA	0.67/1.16	0.44/0.75	0.34/0.55	0.31/0.49	0.50/0.94	0.45/0.77
	SCA	0.66/1.12	0.36/0.58	0.33/0.51	0.29/0.47	0.46/0.85	0.42/0.71
(**Q2**)	d = 8	0.80/1.72	0.33/0.64	0.38/0.84	0.31/0.67	0.50/1.09	0.46/0.99
	d = 12	0.74/1.71	0.30/0.57	0.38/0.86	0.31/0.69	0.50/1.09	0.45/0.98
	d = 16	0.74/1.58	0.30/0.57	0.37/0.81	0.31/0.68	0.50/1.09	0.44/0.95
	d = 24	0.68/1.40	0.29/0.56	0.35/0.79	0.30/0.70	0.48/1.04	0.42/0.90
	d = 32	0.74/1.57	0.31/0.60	0.36/0.86	0.31/0.68	0.50/1.09	0.44/0.96
(**Q3**)	n=8	0.82/1.71	0.31/0.57	0.38/0.83	0.32/0.70	0.50/1.09	0.46/0.98
	n=16	0.77/1.60	0.31/0.60	0.39/0.87	0.30/0.65	0.50/1.09	0.46/0.96
	n=32	0.68/1.40	0.29/0.56	0.35/0.79	0.30/0.70	0.48/1.04	0.42/0.90
	n=64	0.74/1.49	0.30/0.56	0.39/0.86	0.31/0.68	0.50/1.08	0.45/0.93
	n=128	0.76/1.61	0.31/0.56	0.38/0.82	0.31/0.68	0.49/1.08	0.45/0.95
(**Q4**)	$\mathbf{z}$	0.81/1.34	0.77/1.66	0.44/0.83	0.44/0.83	0.75/1.42	0.62/1.19
	$[\mathbf{x}, \mathbf{z}]$	0.98/2.18	1.23/2.61	0.56/1.22	0.48/1.00	0.73/1.62	0.80/1.72
	$\mathbf{x} + \mathbf{z}$	0.68/1.40	0.29/0.56	0.35/0.79	0.30/0.70	0.48/1.04	0.42/0.90

Table 1: Ablation study of our SEAD on the ETH-UCY dataset. Apart from Q1, the ablations are conducted with STGAT. The rows with gray highlight the settings we use. The lower is the better. The standard deviation of our implemented experiments is less than 0.003.

tiveness of SEAD on 4 baseline models and compare them with existing methods.

Experimental Settings

Datasets: Our results are trained on the **ETH** (Pellegrini, Ess, and Gool 2010) and **UCY** (Lerner, Chrysanthou, and Lischinski 2007) datasets. The human trajectories in these datasets are captured in real-world scenes and transformed into sequences of locations. These datasets contain five domains: *ETH, HOTEL, UNIV, ZARA1* and *ZARA2* with a total of 1536 pedestrians detected. All the trajectories are sampled every 0.4 seconds (one frame). We leverage the leave-one-out protocol to split the training, validation and test dataset. Train and validate on four domains and test on the remaining one. We use 3.2 seconds (8 frames) of data as history trajectory and the next 4.8 (12 frames) as the ground truth.

Evaluation Metrics: We use Average Displacement Error (ADE) and Final Displacement Error (FDE) as the evaluation metrics following the previous methods. ADE is the mean square error of the predicted trajectory and ground truth trajectory. FDE is the L2 distance between the final position of the predicted trajectory and the ground truth trajectory:

$$\text{ADE} = \frac{\sum_{i=0}^{m} \sum_{t=T_h+1}^{T_h+T_f} \|\hat{\mathbf{p}}_i^t - \mathbf{p}_i^t\|_2}{m \times T_f} \tag{7}$$

$$\text{FDE} = \frac{\sum_{i=0}^{m} \|\hat{\mathbf{p}}_i^t - \mathbf{p}_i^t\|_2}{m}, t = T_h + T_f. \tag{8}$$

Since the baseline methods, Social-STGCNN, STAGT are generative methods, we follow their evaluation protocol. For each predicted distribution, we sample 20 evaluation trajectories. The one closest to the ground truth trajectory is used for the computation of ADE and FDE.

Implementation Details

As shown in Figure 2, the conventional pipeline of the trajectory prediction models is shown in yellow. Our method shown in green could be integrated into the conventional framework. Generally, SEAD employs the output of the trajectory encoder and leverages causal intervention to remove the negative effect of the social environment. To validate the effectiveness and universality of our proposed SEAD method, we implement our method based on three baseline models, RNN-based STGAT, CNN-based Social-STGCNN and Transformer-based Trajectory Forecasting Transformer (TF). In addition, since our work and Counterfactual analysis(Chen et al. 2021) are complementary, we also conduct experiments with Causal-STGAT with the trajectory predictor $h(\cdot)$ replaced by the counterfactual predictor (in Eq. (6)).

STGAT+SEAD. STGAT is comprised of a M-LSTM, a G-LSTM and a graph attention model (GAT). Specifically, the M-LSTM is used for encoding the trajectory features while GAT and G-LSTM are for encoding social interaction features. Then these features $\mathbf{x}$ are connected and fed into the predictor. To implement our method, we employ these features as the input of our SCA module before they are fed

RNN-based Method	Performance (ADE/FDE)					
	ETH	HOTEL	ZARA1	ZARA2	UNIV	AVG
LSTM	1.09/2.41	0.86/1.91	0.41/0.88	0.52/1.11	0.61/1.31	0.70/1.52
S-LSTM (Alahi et al. 2016)	1.09/2.35	0.79/1.76	0.47/1.00	0.56/1.17	0.67/1.40	0.72/1.54
SGAN (Gupta et al. 2018)	0.81/1.52	0.72/1.61	0.34/0.69	0.42/0.84	0.60/1.26	0.58/1.18
Sophie (Sadeghian et al. 2019)	0.70/1.43	0.76/1.67	0.30/0.63	0.38/0.78	0.54/1.24	0.54/1.15
SR-LSTM (Zhang et al. 2019)	0.63/1.25	0.37/0.74	0.41/0.90	0.32/0.70	0.51/1.10	0.45/0.94
Social-BiGAT (Kosaraju et al. 2019)	0.69/1.29	0.49/1.01	0.30/0.62	0.36/0.75	0.55/1.32	0.48/1.00
MATF (Zhao et al. 2019)	1.33/2.49	0.51/0.95	0.44/0.93	0.34/0.73	0.56/1.19	0.64/1.26
MATF GAN (Zhao et al. 2019)	1.01/1.75	0.43/0.80	0.26/0.45	0.26/0.57	0.44/0.91	0.48/0.90
IDL (Yamaguchi et al. 2011)	0.59/1.30	0.46/0.83	0.22/0.49	0.23/0.55	0.51/1.27	0.40/0.89
PIF (Liang et al. 2019)	0.73/1.65	0.30/0.59	0.60/1.27	0.38/0.81	0.31/0.68	0.46/1.00
STGAT (Huang et al. 2019)	0.65/1.12	0.35/0.66	0.34/0.69	0.29/0.60	0.52/1.10	0.43/0.83
Causal-STGAT (Chen et al. 2021)	0.60/0.98	0.30/0.54	0.32/0.64	0.28/0.58	0.52/1.10	0.40/0.77
STGAT* (Huang et al. 2019)	0.83/1.74	0.31/0.59	0.39/0.85	0.31/**0.68**	0.50/1.08	0.47/0.99
+SEAD	**0.68/1.40**	**0.29/0.56**	**0.35/0.79**	**0.30**/0.70	**0.48/1.04**	**0.42/0.90**
Causal-STGAT* (Chen et al. 2021)	0.69/1.20	**0.31/0.58**	**0.32/0.64**	0.30/0.62	0.54/1.15	0.43/0.86
+SEAD	**0.61/1.08**	0.32/0.60	0.33/**0.64**	**0.27/0.53**	**0.51/1.10**	**0.40/0.79**

Table 2: Comparison with RNN-based methods. The * denotes the results are reproduced with the officially released code. The better result between baseline method and ours is shown in bold. The lower is the better. The standard deviation of our implemented experiments is less than 0.003.

into the predictor. We follow the original implementation details of STGAT. The only difference is that the learning rate for our SCA module is 0.04 while other LSTM and GAT modules keep the same learning rate with STGAT.

Causal-STGAT+SEAD. Similar to STGAT+SEAD, we employ the difference between the factual features and counterfactual features as the input of SCA for Causal-STGAT+SEAD. The learning rate for SCA module in Causal-STGAT is set to 0.01. The hyper-parameters for other modules keep the same with Causal-STGAT.

Social-STGCNN+SEAD. A spatial-temporal graph neural network (STGCNN) is leveraged for trajectory features and social interaction features by Social-STGCNN(Mohamed et al. 2020). To better model the sequential features, time-extrapolator Convolution Neural Networks (TPCNN) are leveraged(Nikhil and Morris 2018). We feed the output of the last TPCNN to our SCA module to adjust the social environment. We follow the original implementations to train Social-STGCNN+SEAD. The initial learning rate is 0.01, and decayed to 0.002 after 150 epochs.

TF+SEAD. Forecasting transformer leverages a transformer encoder to process the observed history positions and a transformer decoder to generate a future trajectory. Our designed SCA module incurs the output of transformer encoder and adjust the social environment. We follow the original implementations to train the whole model.

Ablation Study

Our ablation studies aim to answer the following questions. **Q1**: Does our SEAD module merely benefit from attention mechanism, rather than causal intervention? Is learnable $\mathbf{S}$ indispensable? We demonstrate this by replacing the SCA module with a normal MHSA module and comparing the performance with our method. We conduct such experiments on two baselines: STGAT and Social-STGCNN. **Q2**: What is the proper dimension of $\mathbf{s}$? We conduct experiments with different d. **Q3**: What is the proper number of learnable variables $\mathbf{s}_i$? We conduct experiments with different n. **Q4**: How to make use of the features of SCA module? We experiment on directly feedforward $\mathbf{z}$, concatenation $[\mathbf{x}, \mathbf{z}]$ and residual connection $\mathbf{x} + \mathbf{z}$. Specifically, the number of channels is mapped to 2 after the concatenation for prediction.

the

A1: Results are shown in Table 1 (**Q1**). Compared with baseline models, MHSA module even performs worse. This is because the social interactions have been well learned by the graph neural networks and an extra self attention module could not be beneficial. On both baseline models, our SCA module performs better than the MHSA. Hence, these experimental results show that the superiority of SCA module does not come from the attention mechanism but from causal intervention. The backdoor adjustment is indispensable.

A2: We apply different dimensions d for variable $\mathbf{s}_i$ and modify the dimension of $\mathbf{W}_k, \mathbf{W}_v$ respectively. As shown in Table 1 (**Q2**), d varies from 8 to 32. We could observe that the performance begins to drop when the dimension is larger that 24. In particular, when $d = 24$ our SEAD method achieves the best performance. Therefore, we choose $d = 24$ for the following experiments.

A3: In addition to dimension of $\mathbf{s}_i$, the size of the learnable set $\{\mathbf{s}_i\}$ also matters. The results in Table 1 (**Q3**) shows that when $n = 32$, our SEAD method has the best performance. However, when larger and smaller than 32 the ADE would decrease by $0.03 - 0.04$. This is probably due to that the

CNN/Transformer-based Method	Performance (ADE/FDE)					
	ETH	HOTEL	ZARA1	ZARA2	UNIV	AVG
CNN (Nikhil and Morris 2018)	1.04/2.07	0.59/1.17	0.43/0.90	0.34/0.75	0.57/2.32	0.59/1.22
TF (Giuliari et al. 2021)	1.03/2.10	0.36/0.71	0.44/1.00	0.34/0.76	0.53/1.32	0.54/1.17
Social-STGCNN (Mohamed et al. 2020)	0.64/1.11	0.49/0.85	0.34/0.53	0.30/0.48	0.44/0.79	0.44/0.75
Causal-STGCNN (Chen et al. 2021)	0.64/1.00	0.38/0.45	0.34/0.53	0.32/0.49	0.49/0.81	0.43/0.66
Social-STGCNN* (Mohamed et al. 2020)	0.67/1.13	0.40/0.62	0.34/0.53	0.30/0.48	0.52/0.96	0.45/0.75
+SEAD	**0.66/1.12**	**0.36/0.58**	**0.33/0.51**	**0.29/0.47**	**0.46/0.85**	**0.42/0.71**
TF* (Giuliari et al. 2021)	1.04/2.19	0.43/0.90	0.41/0.92	0.34/0.72	0.59/1.28	0.56/1.20
+SEAD	**1.01/2.07**	**0.35/0.72**	**0.40/0.88**	**0.31/0.68**	**0.57/1.24**	**0.52/1.12**

Table 3: Comparison with CNN/Transformer-based methods. The * denotes the results are reproduced with the officially released code. The better result between baseline method and ours is shown in bold. The lower is the better. The standard deviation of our implemented experiments is less than 0.003.

proper size of the social environment set is 32. A smaller size could not fully model the diversity of the social environment and a larger size would impact the representativeness of the learned variables. We apply $n = 32$ in our experiments.

A4: From Table 1 (**Q4**), we could observe that directly leveraging the output $\mathbf{z}$ and concatenation both perform poor. The residual connection mode has a much better performance than them. The reason for this is probably the optimization complexity of these structures. Therefore, we choose to feed $\mathbf{x} + \mathbf{z}$ into the predictor in our experiments.

Quantitative Analysis

To verify the effectiveness of our method, we reproduce four baseline models and implement our SEAD with their official released code. We compare the performance of the four baseline methods as follows.

Evaluation of RNN-based method. The comparison of our method and RNN-based method and other existing method is summarized in Table 2. Note that results reproduced with officially released code are marked by *. Due to the different implementation environments, the results reported in the original paper are slightly higher than our reproduced results. As shown in Table 2, our implementations could consistently improve the performance of STGAT and Causal-STGAT. Our two implementations improve the ADE/FDE performance by $0.15/0.34$ and $0.08/0.12$ respectively on the *ETH* domain. SEAD has improved the average performance among the 5 domains with $0.05/0.09$ and $0.03/0.08$ respectively.

Besides, we also compare our proposed method with existing *sota* methods, *e.g.*, MATF (Zhao et al. 2019) and IDL (Yamaguchi et al. 2011). As shown in Table 2, our method achieves better performance than IDL on the FDE metric and competitive performance on the ADE metric.

Comparison with Counterfactual Analysis. Even *Counterfactual Analysis* (Chen et al. 2021) incorporates a similar-looking structural causal model, our method and their method are essentially different. They leverage mediation analysis to subtract the natural direct effect of training data while our method incorporates causal intervention to remove the confounding effect of the social environment. These methods are orthogonal. Hence, our method could further improve the performance of Causal-STGAT.

Evaluation of CNN/Transformer-based method. The comparison of our method and CNN/Transformer-based method and other existing method is summarized in Table 3. Our method also achieves consistent improvement over Social-STGCNN. In addition, the ADE and FDE metrics on all the 5 domains have been improved. On the *HOTEL* and *UNIV* domain, our method improves the performance of $0.04/0.04$ and $0.06/0.11$ respectively and achieves an average improvement of $0.03/0.04$ among the 5 domains. Besides, our method outperforms baseline method TF by 0.04/0.08 in total. Our method significantly improves the performances on the *HOTEL* domain by 0.08/0.18.

Conclusion

In this paper, we analyze how the social environment influences human trajectory prediction and we observe that the social environment is a confounder misleading the neural networks to learn spurious correlations between history and future trajectory. Then, we propose a structural causal model to investigate the causalities among social environment, history and future trajectories. Based on causal intervention rather than conventional likelihood, we propose a backdoor adjustment method, named Social Environment Adjustment, to remove the confounding effect according to the proposed structural causal model. We implement the core of our method as a Social Cross Attention module, which is simple yet effective. In addition, the SCA module is universal to improve the performance of various baseline models. Extensive results have demonstrated the effectiveness of our SEAD method. Since we make approximations to implement our method, future work could develop a more advanced confounder representation discovery method.

Acknowledgements

This work is supported in part by the National Science and Technology Major Project of the Ministry of Science and Technology of China under Grant 2018AAA0100701 and the NSFC under Grants 62022048.

References

Alahi, A.; Goel, K.; Ramanathan, V.; Robicquet, A.; Fei-Fei, L.; and Savarese, S. 2016. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 961–971.

Amirian, J.; Zhang, B.; Castro, F. V.; Baldelomar, J. J.; Hayet, J.-B.; and Pettré, J. 2020. Opentraj: Assessing prediction complexity in human trajectories datasets. In *Proceedings of the Asian Conference on Computer Vision*.

Antonini, G.; Bierlaire, M.; and Weber, M. 2006. Discrete choice models of pedestrian walking behavior. *Transportation Research Part B: Methodological*, 40(8): 667–687.

Bai, H.; Cai, S.; Ye, N.; Hsu, D.; and Lee, W. S. 2015. Intention-aware online POMDP planning for autonomous driving in a crowd. In *2015 IEEE international conference on robotics and automation*, 454–460. IEEE.

Chandra, R.; Bhattacharya, U.; Bera, A.; and Manocha, D. 2019. Traphic: Trajectory prediction in dense and heterogeneous traffic using weighted interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8483–8492.

Chen, G.; Li, J.; Lu, J.; and Zhou, J. 2021. Human Trajectory Prediction via Counterfactual Analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9824–9833.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Ge, C.; Huang, R.; Xie, M.; Lai, Z.; Song, S.; Li, S.; and Huang, G. 2022. Domain Adaptation via Prompt Learning. *arXiv preprint arXiv:2202.06687*.

Gheini, M.; Ren, X.; and May, J. 2021. Cross-Attention is All You Need: Adapting Pretrained Transformers for Machine Translation. *arXiv preprint arXiv:2104.08771*.

Giuliari, F.; Hasan, I.; Cristani, M.; and Galasso, F. 2021. Transformer Networks for Trajectory Forecasting. *2020 25th International Conference on Pattern Recognition*, 10335–10342.

Glymour, M.; Pearl, J.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Gupta, A.; Johnson, J.; Fei-Fei, L.; Savarese, S.; and Alahi, A. 2018. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2255–2264.

Hao, Y.; Zhang, Y.; Liu, K.; He, S.; Liu, Z.; Wu, H.; and Zhao, J. 2017. An End-to-End Model for Question Answering over Knowledge Base with Cross-Attention Combining Global Knowledge. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 221–231. Vancouver, Canada: Association for Computational Linguistics.

Helbing, D.; and Molnar, P. 1995. Social force model for pedestrian dynamics. *Physical review E*, 51(5): 4282.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Huang, G.; Ge, C.; Xiong, T.; Song, S.; Yang, L.; Liu, B.; Yin, W.; and Wu, C. 2021. Large scale air pollution prediction with deep convolutional networks. *Science China Information Sciences*, 64(9): 1–11.

Huang, Y.; Bi, H.; Li, Z.; Mao, T.; and Wang, Z. 2019. STGAT: Modeling Spatial-Temporal Interactions for Human Trajectory Prediction. In *The IEEE International Conference on Computer Vision*.

Imbens, G. W.; and Rubin, D. B. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Kosaraju, V.; Sadeghian, A.; Martín-Martín, R.; Reid, I.; Rezatofighi, H.; and Savarese, S. 2019. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *Advances in Neural Information Processing Systems*, 32.

Lerner, A.; Chrysanthou, Y.; and Lischinski, D. 2007. Crowds by example. In *Computer graphics forum*, volume 26, 655–664. Wiley Online Library.

Li, J.; Selvaraju, R. R.; Gotmare, A. D.; Joty, S.; Xiong, C.; and Hoi, S. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *Advances in Neural Information Processing Systems*.

Liang, J.; Jiang, L.; Niebles, J. C.; Hauptmann, A. G.; and Fei-Fei, L. 2019. Peeking into the future: Predicting future person activities and locations in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5725–5734.

Liu, Y.; Cadei, R.; Schweizer, J.; Bahmani, S.; and Alahi, A. 2022. Towards Robust and Adaptive Motion Forecasting: A Causal Representation Perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17081–17092.

Liu, Y.; Yan, Q.; and Alahi, A. 2021. Social nce: Contrastive learning of socially-aware motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15118–15129.

Luber, M.; Stork, J. A.; Tipaldi, G. D.; and Arras, K. O. 2010. People tracking with human motion predictions from social forces. In *2010 IEEE International Conference on Robotics and Automation*, 464–469. IEEE.

Mohamed, A. A.; Qian, K.; Elhoseiny, M.; and Claudel, C. G. 2020. Social-STGCNN: A Social Spatio-Temporal Graph Convolutional Neural Network for Human Trajectory Prediction. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14412–14420.

Morotomi, K.; Katoh, M.; and Hayashi, H. 2014. Collision position predicting device. US Patent 8,849,558.

Musleh, B.; García, F.; Otamendi, J.; Armingol, J. M.; and De la Escalera, A. 2010. Identifying and tracking pedestrians based on sensor fusion and motion stability predictions. *Sensors*, 10: 8028–8053.

Nikhil, N.; and Morris, B. T. 2018. Convolutional Neural Network for Trajectory Prediction. In *Proceedings of the European Conference on Computer Vision Workshops*.

Niu, Y.; Tang, K.; Zhang, H.; Lu, Z.; Hua, X.-S.; and Wen, J.-R. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12700–12710.

Pan, X.; Ge, C.; Lu, R.; Song, S.; Chen, G.; Huang, Z.; and Huang, G. 2022. On the integration of self-attention and convolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 815–825.

Pearl, J. 2009. *Causality*. Cambridge University Press.

Pellegrini, S.; Ess, A.; and Gool, L. V. 2010. Improving data association by joint modeling of pedestrian trajectories and groupings. In *European Conference on Computer Vision*, 452–465. Springer.

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

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Richiardi, L.; Bellocco, R.; and Zugna, D. 2013. Mediation analysis in epidemiology: methods, interpretation and bias. *International journal of epidemiology*, 42(5): 1511–1519.

Sadeghian, A.; Kosaraju, V.; Sadeghian, A.; Hirose, N.; Rezatofighi, H.; and Savarese, S. 2019. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1349–1358.

Tang, K.; Huang, J.; and Zhang, H. 2020. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in Neural Information Processing Systems*, 33: 1513–1524.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Yamaguchi, K.; Berg, A. C.; Ortiz, L. E.; and Berg, T. L. 2011. Who are you with and where are you going? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1345–1352. IEEE.

Zhang, D.; Zhang, H.; Tang, J.; Hua, X.-S.; and Sun, Q. 2020. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33: 655–666.

Zhang, P.; Ouyang, W.; Zhang, P.; Xue, J.; and Zheng, N. 2019. SR-LSTM: State Refinement for LSTM towards Pedestrian Trajectory Prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12085–12094.

Zhao, T.; Xu, Y.; Monfort, M.; Choi, W.; Baker, C.; Zhao, Y.; Wang, Y.; and Wu, Y. N. 2019. Multi-agent tensor fusion for contextual trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12126–12134.