Spatio-Temporal Neural Structural Causal Models for Bike Flow Prediction

Pan Deng[1], Yu Zhao[1*], Junting Liu[1], Xiaofeng Jia[2], Mulan Wang[1]

[1] Beihang University, Beijing, 100191, China.
[2] Beijing Big Data Centre, Beijing, 100024, China.
{pandeng, iyzhao, liujunting, wangmulan}@buaa.edu.cn, jiaxf@jxj.beijing.gov.cn

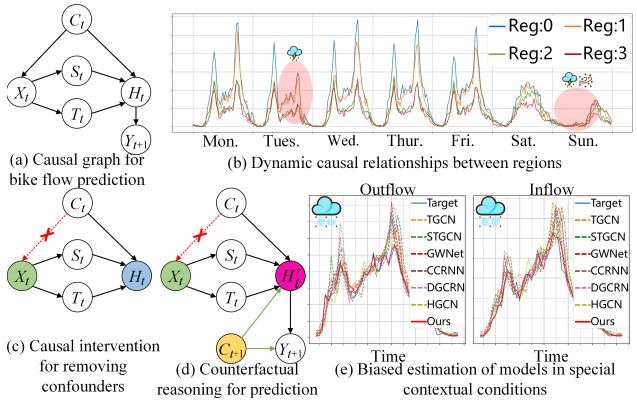
Abstract

As a representative of public transportation, the fundamental issue of managing bike-sharing systems is bike flow prediction. Recent methods overemphasize the spatio-temporal correlations in the data, ignoring the effects of contextual conditions on the transportation system and the inter-regional time-varying causality. In addition, due to the disturbance of incomplete observations in the data, random contextual conditions lead to spurious correlations between data and features, making the prediction of the model ineffective in special scenarios. To overcome this issue, we propose a Spatio-Temporal Neural Structure Causal Model(STNSCM) from the perspective of causality. First, we build a causal graph to describe the traffic prediction, and further analyze the causal relationship between the input data, contextual conditions, spatio-temporal states, and prediction results. Second, we propose to apply the frontdoor criterion to eliminate confounding biases in the feature extraction process. Finally, we propose a counterfactual representation reasoning module to extrapolate the spatio-temporal state under the factual scenario to future counterfactual scenarios to improve the prediction performance. Experiments on real-world datasets demonstrate the superior performance of our model, especially its resistance to fluctuations caused by the external environment. The source code and data will be released.

Introduction

Bike-Sharing systems have been widely deployed in urban public transportation due to their convenience and environmental friendliness in recent years. As a representative of Intelligent Transportation System (ITS), one of the key concerns is the effective allocation of bike-sharing resources to enhance the quality of system service. However, due to the high frequency and randomness of bike usage throughout the city, the stations often get imbalanced over time. The bike-sharing system always has some congested and hungry stations, leading to a significant number of unsatisfied customers. Therefore, it is necessary to accurately predict the bike flow in each commuting area. According to the prediction results of bike flow, it can be repositioned in advance to prevent excessive demand at the station, which is an urgent need for bike-sharing system operators.

*Corresponding author.



(a) Causal graph for bike flow prediction
(b) Dynamic causal relationships between regions
(c) Causal intervention for removing confounders
(d) Counterfactual reasoning for prediction
(e) Biased estimation of models in special contextual conditions

Figure 1: The change process of bike flow from the perspective of causality.

Bike usage patterns and spatio-temporal dependencies between regions are affected by external conditions (e.g. time factors and weather information). As shown in Fig.1(b), the time factors and weather may greatly restrict bike usage in each region, and weekdays and weekends may cause completely different spatio-temporal dependencies. Traditional traffic flow prediction models based on multi-source data usually simply integrate time and weather into the model through the fully connected layer(Li et al. 2019; Liang et al. 2021; Sun et al. 2020). More input information cannot improve the prediction ability of the model. Instead, it will introduce a large number of confounding factors and extract spurious correlations in observations(Liu et al. 2021), resulting in the decline of model performance.

We build a causal graph to describe the bike flow prediction, and further analyze it from a causal perspective. We formulate the causalities among bike flow $\mathbf{X}_t$, contextual condition $\mathbf{C}_t$, spatio-temporal states $\mathbf{H}_t$, and predicted target $\mathbf{Y}_{t+1}$. As shown in Fig. 1(a), directed edges denote causal relationships between nodes. The spatio-temporal state $\mathbf{H}_t$ represents the intrinsic trend of bicycle demand under the current contextual conditions. The contextual condition $\mathbf{C}_t$, as a common cause, affects $\mathbf{X}_t$ and the spatio-temporal state $\mathbf{H}_t$. Bike-sharing systems are susceptible to external contextual conditions, so bike flow prediction must take contextual conditions into account. However, due to the limita-

tion of dataset, the data are collected under normal conditions in most cases, which causes the neural network to only learn the general spatio-temporal patterns under normal conditions, and it is difficult to capture the spatio-temporal state under special conditions. Once a special contextual condition occurs during testing, the predictive performance of the model will drop significantly. As shown in Fig. 1(e), there is a strong snowstorm on this day, the flow is generally low and the rush hours in the morning and evening are not obvious, which is a special environment that unseens in the training set. The predictions of existing methods are on the high side, because they learn a more average spatio-temporal pattern, which is difficult to cope with fluctuations caused by the external environment. These methods emphasize the average performance over the entire dataset while ignoring the prediction performance in specific scenarios. However, bike flow prediction in specific scenarios are often more helpful for managers to formulate emergency measures in advance. Therefore, it is necessary to eliminate the influence of potential contextual conditions on feature extraction by means of causal intervention, so as to make the extracted spatio-temporal state more fair and effectively capture the spatio-temporal pattern of the data itself.

In addition, recent works overemphasize spatio-temporal correlations of bike flow. Generally, these graph-based spatio-temporal prediction models leverage GCN to model spatial correlations, and GRU(Li et al. 2018; Liu et al. 2020; Bai et al. 2020; Ye et al. 2021; Li et al. 2021) or CNN(Wu et al. 2019; Guo et al. 2021a; Fang et al. 2021; Han et al. 2021) to model temporal correlations separately. Although these methods can achieve satisfactory effects, their ability to model complex nonlinear dynamic spatio-temporal causality is still obviously insufficient. As shown in Fig.1(b), bike usage patterns in region 1 are similar to region 2 and 3 during the morning rush hour, but are more similar to region 0 during the evening rush hour. Therefore, the bike flow data implies intense dynamic dependencies in spatial and temporal dimensions and complex nonlinear causality in spatio-temporal dimensions. Most of the adjacency matrices are fixed and generated by heuristic methods based on spatial distance(Liu et al. 2020) or time series similarity(Chai, Wang, and Yang 2018), which cannot capture the time-varying spatio-temporal causality.

To address these aforementioned challenges, We propose a Spatio-Temporal Neural Structural Causal Model (STNSCM), and the causal graph shown in Fig. 1(c) and (d). Its core idea is based on structural causal model theory to remove confounders in the feature extraction process and follow the counterfactual reasoning framework to predict future bike flow. First, we apply the frontdoor criterion based on causal intervention, cutting off the link $\mathbf{C}_t \rightarrow \mathbf{X}_t$, which gives $\mathbf{X}_t$ a fair opportunity to incorporate each contextual condition $\mathbf{C}$ into spatio-temporal state $\mathbf{H}_t$. Second, we view future scenarios in a "what if" way, that is, if the current environment changes, how will the future state change, and thus how will future flow change? The key to answer this counterfactual question is how to make full use of future external conditions. Specifically, The main contributions are as follows:

- We provide a novel causality-based interpretation for the bike flow prediction and apply the frontdoor criterion based on causal interventions to remove confounding biases in the feature extraction process. To the best of our knowledge, our work is the first one that successfully applies the structural causal model to traffic prediction problems.
- We propose a counterfactual representation reasoning module to extrapolate the spatio-temporal state under the factual scenario to the future counterfactual scenario, which enhances the feature's understanding of future states, thereby improving the prediction performance.
- Extensive experiments on two real-world bike-sharing systems datasets show that our STNSCM comprehensively outperforms state-of-the-art methods for region-level bike flow prediction.

Related Work

Spatio-Temporal Traffic Data Prediction

Traffic prediction is a fundamental problem in Intelligent Transportation System including the recent Bike-Sharing System, which has recently attracted significant attention. Since the real-world traffic network is non-Euclidean data, most methods construct GCN models to study spatio-temporal data prediction problems. utilized GCN with predefined graphs to model spatial correlations. To enrich spatial information, (Liu et al. 2020; Fang et al. 2021; Chai, Wang, and Yang 2018) establish multiple static topologies to effectively capture complex patterns. DCRNN(Li et al. 2018) replaces the linear transformation layer in GRU with diffusion graph convolution, which boosts the spatio-temporal representation ability of GRU. HGCN(Guo et al. 2021a) constructs the interaction between the micro and macro layers of GCN, which integrates the different scales of features of road segments and regions.

Recently, some researchers have paid attention to the strong dynamic correlations of traffic data in spatial and temporal dimensions. Therefore, it is crucial to model dynamic and nonlinear spatio-temporal correlations for accurate traffic prediction. GWNet(Wu et al. 2019) and AGCRN(Bai et al. 2020) randomly initialize the node embedding vectors, and then learn an adaptive matrix through gradient descent. Once the training of the model finishes, the adaptive matrix is fixed. So it cannot extract the dynamics from real data effectively. ASTGNN(Guo et al. 2021b) and GMAN(Zheng et al. 2020) dynamically extract the correlations of nodes by means of spatial attention. However, affected by the structure of predefined graphs, spatial attention can only change the weight of predefined graphs rather than the structure. CCRNN(Ye et al. 2021) proposes a coupled graph convolution with self-learned adjacency matrices varying from layer to layer, but the graph structure changes with the convolutional layer instead of time steps. DGCRN(Li et al. 2021) handles the dynamic relations by learning the matrix at each recurrent step, while it significantly depends on the node embedding layer initialized by random parameters. DMSTGCN(Han et al. 2021) learns the time-specific spatial dependencies of road segments to construct dynamic graphs, but

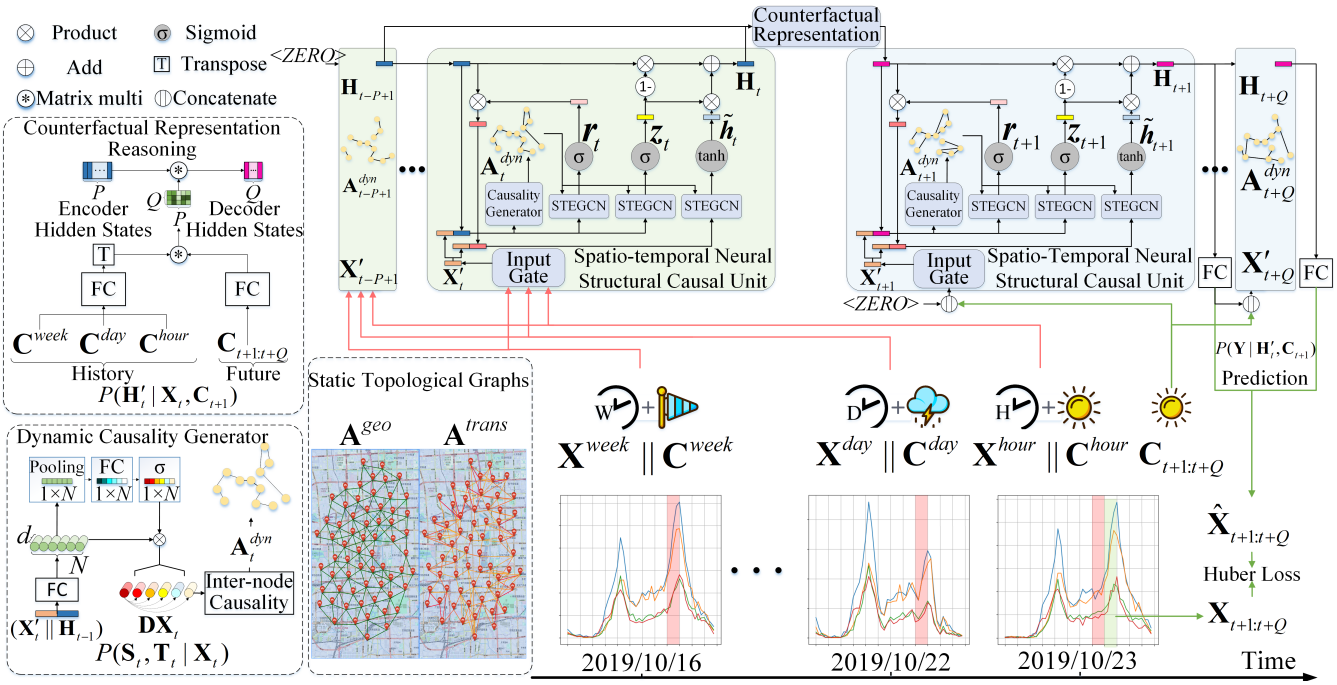


Figure 2: The architecture of STNSCM.

ignores the fluctuations caused by the external environment. In addition, these methods do not take external conditions into consideration during the dynamic graph generation process, resulting in very limited effects of dynamic graphs.

Causal Learning

The purpose of causal learning is to empower models the ability to pursue the causal effect (Zhang et al. 2020). Recent works are proposed to combine SCM (Pearl 2009; Schölkopf 2022) with deep learning models, which is called causal learning. Causal learning is widely used in the field of causal representation, CasualVAE (Yang et al. 2021) proposes a model with causal layer to transform exogenous factors into causal endogenous ones that correspond to causally related concepts in data. Shen (Shen et al. 2020)et al. use a SCM as the prior for bidirectional generative model which can generate data from any desired interventional distributions of the latent factors. (Zhang et al. 2020; Lin et al. 2022; Liu et al. 2022; Yue et al. 2020) use the backdoor criterion to eliminate the contextual confounding biases. Different from above works, the input data and extracted features do not satisfy the backdoor criterion in the field of traffic prediction. As shown in Fig. 1(a), we apply the frontdoor formula and use neural networks to model the sub terms, so our model is called neural structural causal model.

Methodology

The overall framework of STNSCM is shown in Fig.2.

Structural Causal Model

We formulate the causalities among bike flow $\mathbf{X}_t$, contextual condition $\mathbf{C}_t$, spatial neighborhood features $\mathbf{S}_t$, temporal dynamic features $\mathbf{T}_t$, spatio-temporal states $\mathbf{H}_t$, and predicted target $\mathbf{Y}_{t+1}$, with a structural Causal Model (SCM). As shown in Fig. 1(a), directed edges denote causal relationships between nodes. We believe that spatial neighborhood features $\mathbf{S}_t$ and temporal dynamic features $\mathbf{T}_t$ can be decoupled from spatio-temporal data $\mathbf{X}_t$, and they are integrated to form spatiotemporal states $\mathbf{H}_t$, which can describe dynamic spatio-temporal patterns in the data.

Causal Intervention via Frontdoor Criterion

The contextual condition $\mathbf{C}_t$ is expressed as the common cause of $\mathbf{X}_t$ and $\mathbf{H}_t$, which may cause $\mathbf{H}_t$ to be more inclined to the general state and ignore the specific environment due to the limitation of the dataset, resulting in unfair bias of $\mathbf{H}_t$. However, the $\mathbf{C}_t$ have infinite possibilities and cannot be completely covered. we cannot use backdoor adjustments according to the path $\mathbf{X}_t \leftarrow \mathbf{C}_t \rightarrow \mathbf{H}_t$.

Fortunately, we apply the frontdoor criterion based on the path $\mathbf{C}_t \rightarrow \mathbf{X}_t \rightarrow \mathbf{S}_t, \mathbf{T}_t \rightarrow \mathbf{H}_t \leftarrow \mathbf{C}_t$, cutting off the link $\mathbf{C}_t \rightarrow \mathbf{X}_t$, which gives $\mathbf{X}_t$ a fair opportunity to incorporate each contextual condition $\mathbf{C}$ into spatio-temporal state $\mathbf{H}_t$. Formally, we have:

$$
\begin{aligned}
P(\mathbf{H}_t|do(\mathbf{X}_t)) &= \sum_{\mathbf{S}_t,\mathbf{T}_t} P(\mathbf{S}_t,\mathbf{T}_t|do(\mathbf{X}_t))P(\mathbf{H}_t|do(\mathbf{S}_t,\mathbf{T}_t)) \\
&= \sum_{\mathbf{S}_t,\mathbf{T}_t} P(\mathbf{S}_t,\mathbf{T}_t|\mathbf{X}_t) \sum_{\mathbf{X'}_t} P(\mathbf{H}_t|\mathbf{S}_t,\mathbf{T}_t,\mathbf{X'}_t)P(\mathbf{X'}_t)
\end{aligned} \tag{1}
$$

where $P(\mathbf{X'}_t)$ represents the prior distribution of the input data, and we propose an **Input Gate** to fit this prior distribution. $P(\mathbf{S}_t, \mathbf{T}_t|\mathbf{X}_t)$ denotes the process of extracting spatio-temporal features from data and noise, and

we propose a **Dynamic Causality Generator** to embed spatio-temporal causality into a dynamic causal graph. $P(\mathbf{H}_t|\mathbf{S}_t, \mathbf{T}_t, \mathbf{X'}_t)$ denotes the generation of time-varying spatio-temporal states from features to describe the spatio-temporal patterns inherent in the data, and we propose a **Spatio-Temporal Evolutionary Graph Convolution** to extract spatio-temporal states.

Input Gate The periodic flow data are composed of the P slices of bike flow tensor from the previous week $\mathbf{X}^{week}_{t-P+1:t} \in \mathbb{R}^{N\times P\times c_1}$, the previous day $\mathbf{X}^{day}_{t-P+1:t}$, and previous time steps $\mathbf{X}^{hour}_{t-P+1:t}$, where $c_1 = 2$ is the number of flow features. For example, assuming that the time step is one hour, we use the historical four hours ($P = 4$) to predict the bike flow for the next two hours ($Q = 2$) from 18:00-20:00 on Monday. We take the historical periodic flow data as input, including 16:00-20:00 from the previous week, 16:00-20:00 from the previous day and 14:00-18:00 from the previous four hours. Analogously, we can collect the external conditions at the same time step, denoted as $\mathbf{C}^{week}_{t-P+1:t} \in \mathbb{R}^{N\times P\times c_2}$, $\mathbf{C}^{day}_{t-P+1:t}$, $\mathbf{C}^{hour}_{t-P+1:t}$ respectively, where c_2 is the number of external conditions.

As shown in Fig.2, the input gate gets a concatenation of the periodic flow data and the external conditions within the same time step (e.g. $\mathbf{X}^{day}_t||\mathbf{C}^{day}_t$, where $||$ denotes the concatenation operation). The fully connected layer is used to process the input features of week, day, and hour separately. Then we concatenate them together to obtain $\mathbf{X}^{in}_t$. Finally, the gated linear unit is used to output the context-conditioned features $\mathbf{X}'_t$. Given the concatenated input features $\mathbf{X}^{in}_t$, the gated linear unit is defined as follows:

$$\mathbf{X}'_t = \mathbf{X}^{in}_t + \phi(\mathbf{X}^{in}_t\mathbf{\Theta}_1 + \mathbf{a}) \odot \sigma(\mathbf{X}^{in}_t\mathbf{\Theta}_2 + \mathbf{b}) \in \mathbb{R}^{N\times d} \quad (2)$$

where $\mathbf{\Theta}_1$, $\mathbf{\Theta}_2$, $\mathbf{a}$, and $\mathbf{b}$ are model parameters, $\odot$ is the element-wise product, $\phi(\cdot)$ is the tanh function, and $\sigma(\cdot)$ is the sigmoid function.

Dynamic Causality Generator The conditions of bike-sharing systems are complex, and inter-regional relationships are affected not only by spatio-temporal causality, but also by diverse external conditions. However, most methods related to dynamic graphs focus on the spatial correlations between nodes while ignoring the influence of time-varying contextual conditions and causality. As shown in Fig.2, we propose an Dynamic Causality Generator with tightly coupled spatio-temporal states and context-conditioned features.

At each time step, the output of input gate $\mathbf{X}'_t$ and spatio-temporal state of the previous time step $\mathbf{H}_{t-1}$ are concatenated as the input of dynamic causality generator:

$$\mathbf{I}_t = (\mathbf{X}'_t||\mathbf{H}_{t-1})\mathbf{\Theta}_{dyn} + \mathbf{b}_{dyn} \quad (3)$$

where $\mathbf{I}_t \in \mathbb{R}^{N\times d}$, $\mathbf{\Theta}_{dyn}$ and $\mathbf{b}_{dyn}$ are model parameters, d is the number of feature channels. We apply the squeeze-excitation method in the node dimension of $\mathbf{I}_t$ to learn the importance-aware vectorized representation of each node. According to the importance of different nodes, it can promote useful features and suppress features that have little effect on the current task, so that each node can be differentially expressed, which is helpful for subsequent similarity calculation. First, the global channel information is squeezed into the node descriptor using the global average pool:

$$\mathbf{z}_s = \mathrm{F}_{sq}(\mathbf{I}_t) = \frac{1}{d}\sum_{c=1}^{d}\mathbf{I}_t[:, c] \in \mathbb{R}^N \quad (4)$$

Second, the specificity of the node is excited by a self-gating mechanism based on node dependence:

$$\mathbf{z}_e = \mathrm{F}_{ex}(\mathbf{z}_s) = \sigma(\mathbf{\Theta}_{ex2}\,\mathrm{ReLU}(\mathbf{\Theta}_{ex1}\mathbf{z}_s)) \in \mathbb{R}^N \quad (5)$$

where $\mathbf{z}_e \in \mathbb{R}^N$, $\mathbf{\Theta}_{ex1} \in \mathbb{R}^{\frac{N}{16}\times N}$ and $\mathbf{\Theta}_{ex2} \in \mathbb{R}^{N\times\frac{N}{16}}$ are model parameters, $\sigma(\cdot)$ is the sigmoid function, and $\mathrm{ReLU}(\cdot)$ is the ReLU function. Then, $\mathbf{I}_t$ is weighted to generate the dynamic latent representation of each node.

$$\mathbf{DX}_t = \mathrm{F}_{scale}(\mathbf{I}_t, \mathbf{z}_e) = \mathbf{I}_t \odot \mathbf{z}_e \in \mathbb{R}^{N\times d} \quad (6)$$

Finally, we follow the idea of self-attention to calculate the inter-node similarity and embed the dynamic causality into the causal graph:

$$\mathbf{A}^{dyn}_t = \mathrm{ReLU}(\phi(\frac{\mathbf{DX}_t\mathbf{DX}_t{}^T}{\sqrt{d}})) \in \mathbb{R}^{N\times N} \quad (7)$$

where $\phi(\cdot)$ is the tanh function.

Spatio-Temporal Evolutionary Graph Convolution The static topological graphs and the causal graphs based on bike-sharing systems reflect the inter-node relationship from diverse perspectives. Therefore, we propose a Spatio-Temporal Evolutionary Graph Convolution Network (STEGCN) to specify Eq.1, as shown in Fig.2.

For static topological graphs, we establish the geographical distance graph $\mathbf{A}^{geo}$ and the transition probability graph $\mathbf{A}^{trans}$ based on the historical trip records of the regions. Their respective weighted adjacency matrices are as follow:

$$\mathbf{A}^{geo}_{kl} = \begin{cases} \exp(-\frac{dis^2_{kl}}{\sigma^2}) & dis_{kl} > \varepsilon^{geo} \\ 0 & otherwise \end{cases}, \quad \mathbf{A}^{trans}_{kl} = \frac{T_{kl}}{\sum_{c=1}^{N} T_{kc}} \quad (8)$$

where $dis_{kl}, 1 \le k, l \le N$ is distance between region r_k and region r_l calculated by the latitude and longitude of the regional center, ε^{geo} denotes distance threshold and is set as 2 kilometers according to the actual situation, σ is the variance of the distance matrix, which is used to control the distribution and sparsity of matrix, and T_{kl} is transition flow.

The STEGCN can be defined as follows:

$$\begin{aligned} \mathbf{X}^{(0)}_t =& (\mathbf{X}'_t||\mathbf{H}_{t-1}) \in \mathbb{R}^{N\times 2d} \\ \mathbf{X}^{(n)}_t =& \alpha_0\mathbf{X}^{(n-1)}_t + \alpha_1\tilde{\mathbf{A}}^{geo}\mathbf{X}^{(n-1)}_t + \\ & \alpha_2\tilde{\mathbf{A}}^{trans}\mathbf{X}^{(n-1)}_t + \alpha_3\tilde{\mathbf{A}}^{dyn}_t\mathbf{X}^{(n-1)}_t \in \mathbb{R}^{N\times 2d} \\ \tilde{\mathbf{X}}^{out}_t =& \mathrm{ReLU}(\sum_{k=0}^{n}\mathbf{X}^{(k)}_t\mathbf{W}^{(k)} + \mathbf{b}^{(k)}) \in \mathbb{R}^{N\times d} \end{aligned} \quad (9)$$

where n is the depth of propagation, $\mathbf{W}^{(k)}$ and $\mathbf{b}^{(k)}$ are model parameters, α_0, α_1, α_2, α_3 are contribution coefficient that can be learned. The contributions of different

graphs for bike flow prediction can be learned by training α. $\tilde{\mathbf{A}}^{geo}$, $\tilde{\mathbf{A}}^{trans}$ and $\tilde{\mathbf{A}}_t^{dyn}$ denotes the normalized adjacency matrices, defined by $\tilde{\mathbf{A}} = \mathbf{D}^{-1}\mathbf{A}, \mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$. Eq.9 is the specific implementation of Eq.1. The 2nd sum in Eq.1 corresponds to the 2nd formula in Eq.9, indicating that spatio-temporal features are extracted from different aspects (geographic, transition, causality) by means of graph convolution, and they are weighted to sum (i.e. $P(\mathbf{S}_t, \mathbf{T}_t|\mathbf{X}_t) \sum P(\mathbf{H}_t|\mathbf{S}_t, \mathbf{T}_t, \mathbf{X}'_t)P(\mathbf{X}'_t)$). The 1st sum in Eq.1 corresponds to the last formula of Eq.9, indicating that all the above information is combined (i.e. $\sum P(\mathbf{S}_t, \mathbf{T}_t|\mathbf{X}_t) \sum P(\mathbf{H}_t|\mathbf{S}_t, \mathbf{T}_t, \mathbf{X}'_t)P(\mathbf{X}'_t)$).

We simplify Eq.9 to $\mathbf{\Theta} \star_G (\mathbf{X}'_t || \mathbf{H}_{t-1})$. Specifically, we adopt diffusion convolution(Li et al. 2018) to separately propagate the inflow and outflow information of each node in the directed graph, represented as follows:

$$\begin{aligned} \mathbf{X}_t^{out} =& \mathbf{\Theta} \star_G (\mathbf{X}'_t || \mathbf{H}_{t-1}) \\ =& \mathbf{\Theta}_1 \star_G (\mathbf{X}_t^{(0)}, \mathbf{A}) + \mathbf{\Theta}_2 \star_G (\mathbf{X}_t^{(0)}, \mathbf{A}^T) \end{aligned} \tag{10}$$

Spatio-Temporal Neural Structural Causal Unit We integrate the input gate, dynamic causality generator and spatio-temporal evolutionary graph convolution derived from the frontdoor criterion into the Spatio-Temporal Neural Structural Causal Unit (STNSCU) to represent the complete causal intervention process $P(\mathbf{H}_t|do(\mathbf{X}_t))$. As shown in Fig.2, STNSCU can effectively model complex nonlinear spatio-temporal causality, denoted as follows:

$$\begin{aligned} \mathbf{r}_t &= \sigma(\mathbf{\Theta}_r \star_G (\mathbf{X}'_t || \mathbf{H}_{t-1}) + \mathbf{b}_r) \\ \mathbf{z}_t &= \sigma(\mathbf{\Theta}_z \star_G (\mathbf{X}'_t || \mathbf{H}_{t-1}) + \mathbf{b}_z) \\ \tilde{\mathbf{h}}_t &= \phi(\mathbf{\Theta}_h \star_G (\mathbf{X}'_t || (\mathbf{r}_t \odot \mathbf{H}_{t-1})) + \mathbf{b}_h) \\ \mathbf{H}_t &= \mathbf{z}_t \odot \mathbf{H}_{t-1} + (1 - \mathbf{z}_t) \odot \tilde{\mathbf{h}}_t \end{aligned} \tag{11}$$

where $\mathbf{\Theta}_r$, $\mathbf{\Theta}_z$, $\mathbf{\Theta}_h$, $\mathbf{b}_r$, $\mathbf{b}_z$, and $\mathbf{b}_h$ are the parameters of graph convolution, $\star_G$ is graph convolution defined by Eq.10, and $\mathbf{H}_t$ is the spatio-temporal state of the STNSCU at time step t.

In the multi-step forecasting model, STNSCM is an encoder-decoder structure composed of STNSCUs. Historical periodic flow data and external conditions are fed into the encoder and predictions $\hat{\mathbf{X}}^{pred} = \hat{\mathbf{X}}_{t+1:t+Q} \in \mathbb{R}^{N \times Q \times c_1}$ are output by the decoder. The spatio-temporal states of the encoder are transformed to initialize the decoder with a counterfactual representation reasoning module. Since the future external conditions $\mathbf{C}_{t+1:t+Q} \in \mathbb{R}^{N \times Q \times c_2}$ are accessible, it is concatenated with the previous prediction as part of the decoder input.

Counterfactual Representation Reasoning

Under the condition of $\mathbf{X}_t$ in the factual scenario, $\mathbf{C}$ is set to $\mathbf{C}_{t+1}$ to infer the spatio-temporal state $\mathbf{H}'_t$ in the counterfactual scenario, and then predict the future bike flow $\mathbf{Y}_{t+1}$ via $\mathbf{H}'_t$. Formally, we have:

$$\begin{aligned} P(\mathbf{Y}_{C_{t+1}}|\mathbf{X}_t) =& \sum_{\mathbf{H}'_t} P(\mathbf{Y}_{t+1}|do(\mathbf{H}'_t, \mathbf{C}_{t+1}), \mathbf{X}_t) P(\mathbf{H}'_t|\mathbf{X}_t, do(\mathbf{C}_{t+1})) \\ =& \sum_{\mathbf{H}'_t} P(\mathbf{H}'_t \mid \mathbf{X}_t, C_{t+1}) P(\mathbf{Y}_{t+1} \mid \mathbf{H}'_t, \mathbf{C}_{t+1}) \end{aligned} \tag{12}$$

where $P(\mathbf{H}'_t \mid \mathbf{X}_t, \mathbf{C}_{t+1})$ represents the counterfactual representation reasoning process, and $P(\mathbf{Y}_{t+1} \mid \mathbf{H}'_t, \mathbf{C}_{t+1})$ represents the prediction process based on counterfactual representation.

Our purpose is to infer the spatio-temporal state in the case of $\mathbf{C}_{t+1}$ by focusing on the similar part between the external conditions of the future and history. Counterfactual representation reasoning module employs scaled dot-product attention to dynamically calculate relationships between each future and historical time step, and converts the encoded historical spatio-temporal states to future representations. Finally, future representations are used to initialize the decoder. As shown in Fig.2, a fully connected layer is used to generate future external features $\mathbf{F}^{pred} = \mathrm{FC}(\mathbf{C}_{t+1:t+Q}) \in \mathbb{R}^{N \times Q \times d}$ and historical external features $\mathbf{C}^{his} = \mathrm{FC}(\mathbf{C}^{week} || \mathbf{C}^{day} || \mathbf{C}^{hour}) \in \mathbb{R}^{N \times P \times d}$, which are taken as the query and key of the attention mechanism. The historical spatio-temporal states $\mathbf{H}^{his} = \mathbf{H}_{t-P+1:t} \in \mathbb{R}^{N \times P \times d}$ are taken as the value. The counterfactual representation reasoning module is formulated as:

$$\begin{aligned} \mathbf{Q} &= \mathbf{C}^{pred}\mathbf{W}^Q \in \mathbb{R}^{N \times Q \times d} \\ \mathbf{K} &= \mathbf{C}^{his}\mathbf{W}^K \in \mathbb{R}^{N \times P \times d} \\ \mathbf{V} &= \mathbf{H}^{his}\mathbf{W}^V \in \mathbb{R}^{N \times P \times d} \\ \mathbf{H}^{pred} &= \mathrm{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V} \in \mathbb{R}^{N \times Q \times d} \end{aligned} \tag{13}$$

where d is the number of feature channels. $\mathbf{W}^Q$, $\mathbf{W}^K$, and $\mathbf{W}^V$ are learnable parameters. Intuitively, for historical spatio-temporal states, the counterfactual representation reasoning module indicates to pay more attention to the parts whose external conditions are similar to the future. The future representations $\mathbf{H}^{pred}$ are input into a fully connected layer and then used to initialize the decoder.

Experiments

In this section, we evaluate the effectiveness of STNSCM by experiments conducted on real-world datasets[1].

Experimental Settings

Datasets: We collect two real-world datasets, NYC-Bike and BJ-Bike, each dataset contains the corresponding weather and time information. We split the dataset with a 30-minute interval. We select the first 60% of data as the training set, 20% as the validation set, and 20% as the test set.

Baselines: We compare STNSCM with recent state-of-the-art baselines. We summarize the models into four categories, including deep learning methods, predefined graph methods, adaptive graph methods, attention methods and dynamic graph methods. The main difference between adaptive graph and dynamic graph is whether the change of graph structure depends on input features.

Evaluation Metrics: We evaluate the performance of methods with Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE).

[1] https://github.com/EternityZY/STNSCM

	Category	Models	30min			60min			Avg		
			MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
BJ-Bike	Deep Learning	LSTM	14.203	28.216	19.530%	19.890	41.400	26.170%	17.37	35.313	22.982%
		GRU	14.506	30.584	20.177%	20.289	42.866	27.120%	17.729	37.076	23.791%
	Predefined Graph	STGCN	11.522	32.306	15.427%	14.158	36.154	18.361%	13.201	33.434	16.986%
		STGODE	13.072	25.508	17.693%	18.286	38.222	23.712%	15.974	32.067	20.828%
	Adaptive Graph	CCRNN	13.002	40.762	16.266%	15.160	43.877	19.046%	14.429	40.767	17.753%
		DMSTGCN	11.396	23.109	15.662%	14.128	29.565	18.900%	12.9 92	25.664	17.352%
	Attention Graph	GMAN	14.297	32.540	19.485%	19.541	43.854	25.745%	17.306	38.488	22.745%
		ASTGNN	13.049	26.241	17.426%	17.831	40.430	23.151%	15.810	33.464	20.427%
	Dynamic Graph	DGCRN	11.416	27.705	16.087%	14.046	33.232	19.200%	12.996	29.817	17.758%
		STNSCM	**11.283**	**23.289**	**15.297%**	**13.341**	**28.125**	**17.472%**	**12.518**	**24.438**	**16.487%**
NYC-Bike	Deep Learning	LSTM	3.125	6.059	24.842%	3.834	7.911	30.982%	3.480	6.904	28.599%
		GRU	3.135	6.062	24.876%	3.844	7.869	30.940%	3.490	6.882	28.614%
	Predefined Graph	STGCN	2.601	4.907	20.506%	2.973	6.107	23.367%	2.787	5.404	22.465%
		STGODE	2.722	5.239	21.448%	3.207	6.711	25.468%	2.964	5.840	23.894%
	Adaptive Graph	CCRNN	2.594	4.898	20.414%	2.967	6.243	23.462%	2.781	5.488	22.495%
		DMSTGCN	2.553	4.740	20.186%	2.915	6.037	23.215%	2.739	5.320	22.147%
	Attention Graph	GMAN	3.115	6.264	23.675%	3.181	6.459	24.744%	3.146	6.164	24.713%
		ASTGNN	2.977	5.656	23.384%	3.334	6.828	26.281%	3.157	6.023	25.360%
	Dynamic Graph	DGCRN	2.617	4.942	20.583%	2.965	6.141	23.342%	2.791	5.421	22.485%
		STNSCM	**2.528**	**4.683**	**19.847%**	**2.809**	**5.612**	**22.445%**	**2.670**	**5.039**	**21.530%**

Table 1: Performance comparison with other models.

Comparison with Baselines

For fairness, we deploy the same environment, loss function, periodic flow data, and external conditions for all models. We compare STNSCM with baselines for traffic prediction and the final average results are shown in Table 1.

We classify these methods according to whether the graph structure varies depending on input features. Results show our STNSCM outperforms baseline models consistently and overwhelmingly. In deep learning methods, Poor performances of indicate the limitation of failing to consider spatial correlation. STGODE(Fang et al. 2021) deepens networks to extract higher-order features through a tensor-based ordinary differential equation, but hampered by the amount of data, STGODE shows a worse performance. Besides, these methods only rely on the fixed graph structure while ignoring the dynamic characteristics.

The adaptive graph(Ye et al. 2021; Han et al. 2021) is helpful in the short-term prediction (30min), but it is still static over time and fails to capture time-varying spatio-temporal dependencies, and the effect of long-term prediction (60min) is significantly reduced. The attention graph(Zheng et al. 2020; Guo et al. 2021b) can only change the weight of predefined graphs rather than the structure. DGCRN(Li et al. 2021) employs the pre-defined adjacency matrix to conduct the message-passing process for dynamic node status. Due to the incomplete connections in the data, the predefined graph itself may contain noise, hindering the generation of dynamic graphs.

Besides, these models do not have exclusive modules to handle contextual conditions, and we merely concatenate them with periodic flow data, so external information is not fully exploited. The qualitative prediction results are shown in Fig. 1(e). The main contribution of our model is stability and resistance to random fluctuations, which are also the most important capabilities of bike flow forecasting. Therefore, in the NYC-Bike test set, the bike flow distribution fluctuates greatly due to the external conditions, so our method outperforms all methods.

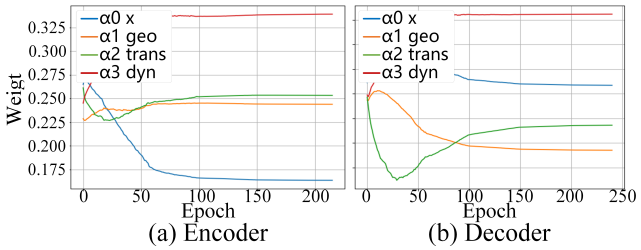
Category	Models	Average		
		MAE	RMSE	MAPE
Graph	w/o $\mathbf{A}^{geo}$	2.710	5.218	21.845%
	w/o $\mathbf{A}^{trans}$	2.712	5.251	22.078%
	w/o $\mathbf{A}^{dyn}$	2.741	5.314	22.331%
Dynamic Causality Generator	EGG w/o SE	2.707	5.192	21.882%
	EGG w/o H	2.730	5.313	22.293%
	EGG w/o X	2.710	5.261	22.025%
Input Gate	IG w/ FC	2.726	5.305	22.224%
	w/o IG	2.723	5.335	22.317%
Counterfactual Representation	w/o CR	2.704	5.209	21.843%
	STNSCM	**2.670**	**5.039**	**21.530%**

Table 2: Comparison with variants on NYC-Bike.

Component Analysis

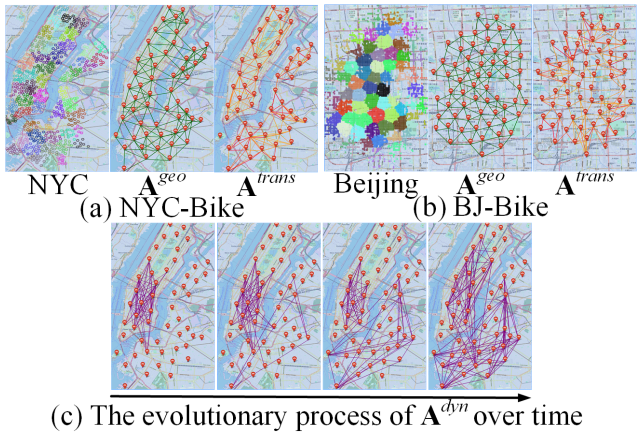
To verify effectiveness of key components in STNSCM, we conduct ablation experiments on NYC-Bike dataset, which are described as follows:

- w/o $\mathbf{A}^{geo}$: It removes the geographical distance graph.
- w/o $\mathbf{A}^{trans}$: It removes the transition probability graph.
- w/o $\mathbf{A}^{dyn}$: It removes the dynamic causal graph.
- EGG w/o SE: It removes the squeeze-excitation method from dynamic causality generator.



(a) Encoder (b) Decoder

Figure 3: The change process of the contribution coefficient α during the training period.



(a) NYC-Bike (b) BJ-Bike

(c) The evolutionary process of $\mathbf{A}^{dyn}$ over time

Figure 4: Static topologies and dynamic causal graphs.

- EGG w/o $\mathbf{H}$: It removes the spital-temporal state of the previous time step from dynamic causality generator.
- EGG w/o $\mathbf{X}$: It removes the features output by input gate from dynamic causality generator.
- w/o CR: It removes the counterfactual representation reasoning module from STNSCM. The last spital-temporal state of the encoder are copied to initialize the decoder.
- IG w/ FC: It concatenates periodic flow data and external conditions to extract context-conditioned features only through fully connected layers.
- w/o IG: It removes the input gate. The model merely input the bike flow tensor of the previous P time steps.

The performance of all variants is summarized in Table 2. For the contribution of different graphs, the dynamic causal graph obviously plays a more prominent role. The introduction of the causal graph can significantly improve performance, as it provides implicit causality that cannot be extracted from static topological graphs. Meanwhile, the geographical distance graph and the transition probability graph are also necessary. The dynamic causal graph can collaborate with static topological graphs to better model complex transportation systems. We further visualized the change process of the contribution coefficient α in the spatio-temporal evolutionary graph convolution during the training period, as shown in Fig.3. In the encoder, the geographic distance graph and the transition probability graph have the same importance, while the input feature $\mathbf{X}_t^{(n-1)}$ of the graph convolution has a smaller coefficient. This indicates that the encoder needs to extract potential spatio-temporal dependencies in the transportation system by propagating and aggregating the node features from multi-view graphs. On the contrary, in the decoder, the input features $\mathbf{X}_t^{(n-1)}$ of the graph convolution account for a large proportion, which shows that the decoder needs to restore high-level spatio-temporal features to predict the future flow.

For the dynamic causality generator, the spatio-temporal state of the previous time step is crucial for modeling the dynamic spatio-temporal causality into the causal graph. We visualized the dynamic causal graph, as shown in Fig.4(c), the dynamic causal graph integrated with periodic features and external conditions directly model the interactive evolutionary process across regions.

In addition, external conditions dominate the model performance, otherwise, the model cannot govern the fluctuations caused by this factor. On the one hand, the performance of IG w/ FC and w/o IG is almost the same. This indicates that features extracted by our input gate are more effective. On the other hand, the model's resistance to random fluctuations can be reflected in MAPE. Removing the input gate will greatly increase MAPE. The input gate can effectively fuse periodic flow data and external conditions to extract the context-conditioned features, which is also the basis for STNSCM to resist fluctuations caused by external environment.

The counterfactual representation reasoning module utilizes the attention mechanism to convert the encoded historical spatio-temporal states to future representations, which reduces information loss and improves the overall performance of the model.

Conclusion

In this work, we build a causal graph to describe the traffic prediction problem from a perspective of causality. It shows that due to the disturbance of incomplete observation, there are spurious correlations in the feature extraction process, resulting in the model can only perform general scenarios, but failing in special scenarios. We propose a novel spatio-temporal neural structural causal model that decomposes the frontdoor criterion into multiple sub-terms and proposes well-designed modules to model these subterms. Among them, the dynamic causality generator is the most important. It embeds the inter-regional time-varying causal relationship into the dynamic causal graph, which enables the model to capture the dynamic rules. Second, we propose a counterfactual representation reasoning module, which makes spatio-temporal states in the current factual scenarios have the ability to represent counterfactuals. Detailed experiments and analyses demonstrated the superiority of STNSCM over both classical and state-of-the-art prediction methods.

References

Bai, L.; Yao, L.; Li, C.; Wang, X.; and Wang, C. 2020. Adaptive Graph Convolutional Recurrent Network for Traf-

fic Forecasting. In *34th Conference on Neural Information Processing Systems*.

Chai, D.; Wang, L.; and Yang, Q. 2018. Bike Flow Prediction with Multi-Graph Convolutional Networks. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 397–400. New York, NY, USA: Association for Computing Machinery. ISBN 9781450358897.

Fang, Z.; Long, Q.; Song, G.; and Xie, K. 2021. Spatial-temporal graph ode networks for traffic flow forecasting. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 364–373.

Guo, K.; Hu, Y.; Sun, Y.; Qian, S.; Gao, J.; and Yin, B. 2021a. Hierarchical Graph Convolution Networks for Traffic Forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 151–159.

Guo, S.; Lin, Y.; Wan, H.; Li, X.; and Cong, G. 2021b. Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting. *IEEE Transactions on Knowledge and Data Engineering*.

Han, L.; Du, B.; Sun, L.; Fu, Y.; Lv, Y.; and Xiong, H. 2021. Dynamic and Multi-faceted Spatio-temporal Deep Learning for Traffic Speed Forecasting. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 547–555.

Li, F.; Feng, J.; Yan, H.; Jin, G.; Jin, D.; and Li, Y. 2021. Dynamic Graph Convolutional Recurrent Network for Traffic Prediction: Benchmark and Solution. *arXiv preprint arXiv:2104.14917*.

Li, Y.; Yu, R.; Shahabi, C.; and Liu, Y. 2018. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In *International Conference on Learning Representations (ICLR '18)*.

Li, Y.; Zhu, Z.; Kong, D.; Xu, M.; and Zhao, Y. 2019. Learning heterogeneous spatial-temporal representation for bike-sharing demand prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 1004–1011.

Liang, Y.; Ouyang, K.; Sun, J.; Wang, Y.; Zhang, J.; Zheng, Y.; Rosenblum, D.; and Zimmermann, R. 2021. Fine-Grained Urban Flow Prediction. In *Proceedings of the Web Conference 2021*, 1833–1845.

Lin, X.; Chen, Y.; Li, G.; and Yu, Y. 2022. A Causal Inference Look at Unsupervised Video Anomaly Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2): 1620–1629.

Liu, C.; Sun, X.; Wang, J.; Tang, H.; Li, T.; Qin, T.; Chen, W.; and Liu, T.-Y. 2021. Learning causal semantic representation for out-of-distribution prediction. *Advances in Neural Information Processing Systems*, 34.

Liu, L.; Chen, J.; Wu, H.; Zhen, J.; Li, G.; and Lin, L. 2020. Physical-virtual collaboration modeling for intra-and inter-station metro ridership prediction. *IEEE Transactions on Intelligent Transportation Systems*.

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.

Pearl, J. 2009. *Causality*. Cambridge university press.

Schölkopf, B. 2022. Causality for machine learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, 765–804.

Shen, X.; Liu, F.; Dong, H.; Lian, Q.; Chen, Z.; and Zhang, T. 2020. Disentangled generative causal representation learning. *arXiv preprint arXiv:2010.02637*.

Sun, J.; Zhang, J.; Li, Q.; Yi, X.; Liang, Y.; and Zheng, Y. 2020. Predicting citywide crowd flows in irregular regions using multi-view graph convolutional networks. *IEEE Transactions on Knowledge and Data Engineering*.

Wu, Z.; Pan, S.; Long, G.; Jiang, J.; and Zhang, C. 2019. Graph WaveNet for Deep Spatial-Temporal Graph Modeling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, 1907–1913.

Yang, M.; Liu, F.; Chen, Z.; Shen, X.; Hao, J.; and Wang, J. 2021. CausalVAE: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9593–9602.

Ye, J.; Sun, L.; Du, B.; Fu, Y.; and Xiong, H. 2021. Coupled Layer-wise Graph Convolution for Transportation Demand Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 4617–4625.

Yue, Z.; Zhang, H.; Sun, Q.; and Hua, X.-S. 2020. Interventional few-shot learning. *Advances in neural information processing systems*, 33: 2734–2746.

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.

Zheng, C.; Fan, X.; Wang, C.; and Qi, J. 2020. Gman: A graph multi-attention network for traffic prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 1234–1241.