Causal Conditional Hidden Markov Model for Multimodal Traffic Prediction

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

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

Abstract

Multimodal traffic flow can reflect the health of the transportation system, and its prediction is crucial to urban traffic management. Recent works overemphasize spatio-temporal correlations of traffic flow, ignoring the physical concepts that lead to the generation of observations and their causal relationship. Spatio-temporal correlations are considered unstable under the influence of different conditions, and spurious correlations may exist in observations. In this paper, we analyze the physical concepts affecting the generation of multimode traffic flow from the perspective of the observation generation principle and propose a Causal Conditional Hidden Markov Model (CCHMM) to predict multimodal traffic flow. In the latent variables inference stage, a posterior network disentangles the causal representations of the concepts of interest from conditional information and observations, and a causal propagation module mines their causal relationship. In the data generation stage, a prior network samples the causal latent variables from the prior distribution and feeds them into the generator to generate multimodal traffic flow. We use a mutually supervised training method for the prior and posterior to enhance the identifiability of the model. Experiments on real-world datasets show that CCHMM can effectively disentangle causal representations of concepts of interest and identify causality, and accurately predict multimodal traffic flow.

Introduction

Urban transportation systems are generally multimodal in nature, consisting of several interconnected subsystems representing different modes of transportation, such as bike, taxi, bus and car. They aim to meet diverse travel demands and provide residents with a variety of travel options(Liang, Huang, and Zhao 2021). Multimodal traffic flow can reflect the health of the transportation system. Urban traffic managers can formulate corresponding management strategies according to the traffic flow in different environments to improve the smoothness of urban operation. Therefore, multimodal traffic flow prediction is a key part of urban traffic management, providing important data support for traffic guidance (Liang, Huang, and Zhao 2021).

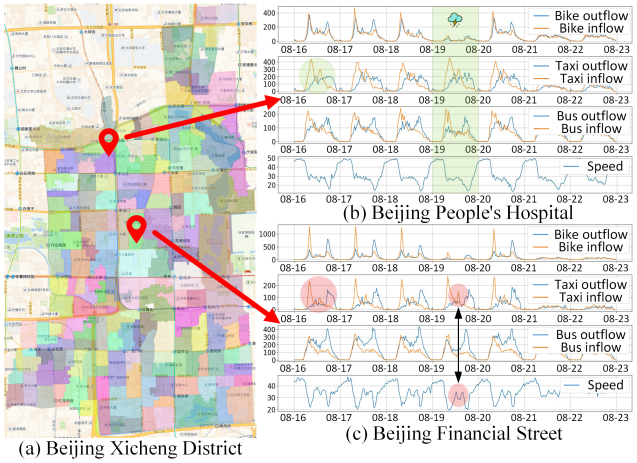

Figure 1: Multimodal traffic flow in different regions.

Most methods only predict a certain traffic flow (e.g., taxi demand or speed)(Bai et al. 2020; Li et al. 2021b; Wu et al. 2020; Ye et al. 2021; Han et al. 2021). They are only partial observations of the traffic system and cannot truly reflect the real situation in real-world scenarios. In contrast, the existing multimodal traffic prediction methods often take different traffic flows as the channel expansion of input data(Wang et al. 2021; Li et al. 2021a; Zhou et al. 2021; Liang et al. 2021), or integrate the feature representation of different flows in the model(Ye et al. 2019; Deng et al. 2021). They implicitly extract the so-called spatio-temporal correlations while lacking the description of causality. However, 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(Schölkopf et al. 2021; Liu et al. 2021a; Deng and Zhang 2021), resulting in the decline of model performance.

Nowadays, traffic flow prediction method overemphasizes spatio-temporal correlations of traffic flow(Liu et al. 2021b; Bai et al. 2020; Li et al. 2021b; Ye et al. 2021; Han et al. 2021), ignoring the physical concepts that lead to the generation of observations and the causal relationship between these concepts. Spatio-temporal correlations are considered unstable under the influence of different conditions, and spu-

*Corresponding author.

rious correlations may exist in observations. Causality is necessary when we delve into the generation principle of observation. For example, researchers(Ye et al. 2019; Deng et al. 2021) believe that there is a certain correlation between taxi and bike flow, and that they can boost each other in terms of multi-task learning. As shown in Fig. 1(b), the flow of taxis and bikes seems to be correlated under normal conditions. Since people's demand for arriving or leaving a region is consistent during rush hours, the tends are similar. However, when it rains (marked in red), the demand for bikes decreases due to weather changes, but the demand for taxis increases, with diametrically opposite trends during the same period. This indicates a spurious correlation between taxis and bike flow under the influence of weather. Second, we believe that the regional attribute has a strong causal relationship with people's travel demand. As shown in Fig. 1(b) and (c), the area has a strong regional attraction under the influence of the hospital attribute, leading to a large demand for people, so it has obvious rush hours in morning and noon. In addition, this area is chronically congested due to high demand(marked in green). Beijing Financial Street is the main working area with a large number of enterprises, so it has obvious morning and evening rush hours (marked in green). We provide more examples of regional POI elements affecting travel demand in appendix. Finally, the demand for taxis may have an impact on the traffic speed. As shown in Fig.1(c). The taxi demand can be inferred from flow. The larger the taxi flow, the more vehicles on the road, and the slower traffic speed (marked in blue). By contrast, High bus demand does not mean a large number of buses on the road, so there is little causal relationship between bus demand and speed.

According to the above analysis, the essential factor affecting multimodal traffic observations is the causal relationship with physical concepts, and excessive attention to the correlation will lead to unstable prediction results. we rethink the generation process of multimodal traffic flow, and explicitly separate the core physical concepts affecting the observation generation into three groups: 1) The attraction factor of the region to people in different time periods. 2) The demand factor (including bikes, taxis and buses) of people choosing different transportation modes under different conditions, and 3) The speed factor affected by the number of vehicles on the road. Our primary task is to disentangle the causal representation of these concepts from conditional information and observations, and further explore their causal relationship.

In this paper, we regard the spatio-temporal multimodal traffic sequence generation process as a Conditional Markov Process, and propose a Causal Conditional Hidden Markov Model (CCHMM). We disentangle the underlying explanatory factors by means of Variational inference, and establish the causal relationship between latent variables by using the Structural Causal Model (SCM)(Pearl 2009; Schölkopf 2022). Compared with the existing work, instead of building a complex adjacency graph between regions to extract the spatio-temporal correlations in the observation data, we model multimodal traffic flow prediction from a causal perspective. The theoretical innovation in the field of traffic forecasting is as follows: Based on the idea of causality, we model the operation process of multimodal traffic systems from the perspective of the observation generation principle, while the existing methods do not focus on causality in the observation data. We propose a causal graph (shown in Fig. 2) to describe the operation of multimodal traffic systems, on which we define a joint distribution (shown in Eq. 1) that describes the principle of observation data generation. Specifically, first, the posterior network infers the disentangled representation of concepts of interest from conditional information and observation data and learns the variational posterior distribution. Then, the prior network models the natural physical laws that existed in the system from the conditional information and learns the prior distribution of the concepts of interest. Third, the causal propagation module mines the causal effects and transforms the exogenous variables inferred from the prior and posterior networks into causal endogenous variables. Finally, The causal endogenous variables are fed into the generator to generate multimodal traffic flow and regarded as the prediction results. The main contributions of this work are as follows:

- We analyze the core physical concepts that affect the multimodal traffic flow generation process, disentangle the causal representations of concepts of interest, and further explore their causal relationship.
- We reform the previous prediction methods and innovatively propose a Causal Conditional Hidden Markov Model (CCHMM) to predict multimodal traffic flow from the perspective of observation generation principle.
- We propose a mutually supervised training method for the prior and posterior to capture physical rules of concepts and enhance the causal identifiability of the model.
- Extensive experiments on real-world datasets show that CCHMM comprehensively outperforms state-of-the-art methods for multimodal traffic flow prediction.

Related Works

Multimodal Traffic Flow Prediction. Giving the increasing availability of diverse data sources, most recent studies has focused on the multimodal fusion in traffic flow prediction. Researchers construct models based on multi-task learning framework to forecast traffic flow and speed simultaneously(Wang et al. 2021; Li et al. 2021a). Ye (Ye et al. 2019) et al. decompose spatial traffic flow with a convolutional autoencoder and implement heterogeneous LSTM for predicting traffic flow of three traffic modes simultaneously. Deng (Deng et al. 2021) et al. learn multi-view representations for single-modal traffic flow and introduce a cross-view self-attention mechanism to capture the co-evolution correlation between different traffic modes. Most of these works implemented Multilayer Perception (MLP) for encoding conditional information(e.g. weather and POI) utilized CNN(Liang et al. 2021; Cao et al. 2021) or Graph Convolutional networks (GCN)(Wu et al. 2019; Han et al. 2021) for capturing spatial features and used RNN for temporal features(Ye et al. 2021; Li et al. 2021b; Bai et al. 2020). Finally, the fused features are fed into downstream prediction network. However, these models do not distinguish the

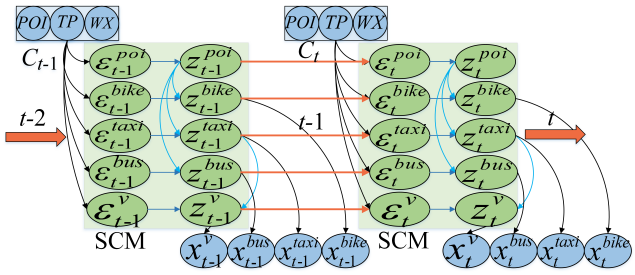

Figure 2: The causal graph of multimodal transportation systems

features related to different tasks, which make models learn spurious correlations during the training process. The spurious correlations make models difficult to generalize beyond their training distribution.

Causal Disentangled Representation Learning. In representation learning, the observation x is generated by a two-step generative process. First, the latent variable z is sampled from a prior distribution $p(z)$, and then the observation x is sampled from the conditional distribution $p(x|z)$(Locatello et al. 2019). Disentangled representation learning aims to learn separable latent variables $z = \{z_1, z_2, \ldots, z_n\}$. Most existing methods rely on the independency assumption of latent variables which is potentially unrealistic(Khemakhem et al. 2020). In fact, there is generally a complex causal relationship between latent variables(Yang et al. 2021). To address this issue, recent works are proposed to combine SCM (Pearl 2009; Schölkopf 2022) with deep learning models. 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. Different from above works, our model focused on causal disentangled representation learning on spatial-temporal series. Li (Li et al. 2021c) et al. propose a time series disease forecasting method based on HMM. Although this method can disentangle the latent variables that are related to disease, while ignores the casual relationship among factors. In our model, we construct a comprehensive temporal causal graph for conditional information, latent variables and observation data. To the best of our knowledge, our work is the first one that successfully applies the structural causal model to traffic prediction problems.

Methodology

Problem Definition

We define the generation process of multimodal traffic flow as a Conditional Markov Process, illustrated as a Directed Acyclic Graph(DAG), as shown in Fig 2. For the latent variable inference stage at time step t, the conditional information $\mathbf{C}_t$ composed of POI, time position TP_t and weather WX_t reflects the current system external status. The conditional information is combined with causal endogenous latent variables $\mathbf{z}_{t-1}$ from the previous time step $t-1$ to extract independent exogenous variables $\boldsymbol{\epsilon}_t = \left[\boldsymbol{\epsilon}_t^{poi}, \boldsymbol{\epsilon}_t^{bike}, \boldsymbol{\epsilon}_t^{taxi}, \boldsymbol{\epsilon}_t^{bus}, \boldsymbol{\epsilon}_t^{v}\right]$, which is determined by the system external status and are not affected by observations. Then, the Structural Causal Model (SCM) $\mathbf{z}_i \leftarrow f(pa(\mathbf{z}_i), \boldsymbol{\epsilon}_i)$(Schölkopf 2022) assigns the generative mechanism of each latent endogenous variable, where $pa(\mathbf{z}_i)$ denotes the set of parent nodes of $\mathbf{z}_i$. It transforms the independent exogenous variables to causal endogenous variables $\mathbf{z}_t = \left[\mathbf{z}_t^{poi}, \mathbf{z}_t^{bike}, \mathbf{z}_t^{taxi}, \mathbf{z}_t^{bus}, \mathbf{z}_t^{v}\right]$. The causal endogenous latent variables $\mathbf{z}_t$ are regarded as an approximate representation of a series of concepts of interest, where the elements represent the regional attraction factor, bike demand factor, taxi demand factor, bus demand factor and speed factor at time t, respectively. Since these latent variables evolve as intrinsic drivers for the progression of multimodal traffic observations, the prior distribution of the latent variables has Markov property and is defined as $p(\boldsymbol{\epsilon}_t, \mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{C}_t) = p(\boldsymbol{\epsilon}_t|\mathbf{z}_{t-1}, \mathbf{C}_t) * p(\mathbf{z}_t|\boldsymbol{\epsilon}_t)$.

For the data generation stage at time step t, the exogenous latent variables $\boldsymbol{\epsilon}_t$ are sampled from the prior distribution $p(\boldsymbol{\epsilon}_t|\mathbf{z}_{t-1}, \mathbf{C}_t)$, The causal endogenous latent variables $\mathbf{z}_t$ are generated using a SCM. Finally, the observations $\mathbf{x}_t = \left[\mathbf{x}_t^{bike}, \mathbf{x}_t^{taxi}, \mathbf{x}_t^{bus}, \mathbf{x}_t^{v}\right]$ are generated from the conditional distribution $p(\mathbf{x}_t|\mathbf{z}_t)$.

A Probabilistic Generative Model for CCHMM

We give the joint distribution definition of the probabilistic generative model of CCHMM and factorize it according to the DAG (Fig. 2) and Causal Markov Condition(Pearl 2009):

$$\begin{aligned} &p_\theta\left(\mathbf{x}_{t<T}, \boldsymbol{\epsilon}_{t<T}, \mathbf{z}_{t<T} \mid \mathbf{C}_{t<T}\right) \\ &= \prod_{t=1}^{T-1} p_\theta\left(\boldsymbol{\epsilon}_t, \mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{C}_t\right) * p_\theta\left(\mathbf{x}_t \mid \mathbf{z}_t\right) \end{aligned} \tag{1}$$

The first term is the prior model, which can be further factored into the generative mechanism of exogenous and endogenous variables based on the causal relationship:

$$p_\theta\left(\boldsymbol{\epsilon}_t, \mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{C}_t\right) = p_\theta\left(\boldsymbol{\epsilon}_t \mid \mathbf{z}_{t-1}, \mathbf{C}_t\right) * p_\theta\left(\mathbf{z}_t \mid \boldsymbol{\epsilon}_t\right) \tag{2}$$

The second item is the generative model, which can be further factored into generative models for each modality depending on endogenous variables corresponding to concepts of interest:

$$\begin{aligned} p_\theta\left(\mathbf{x}_t \mid \mathbf{z}_t\right) = p_\theta\left(\mathbf{x}_t^{bike} \mid \mathbf{z}_t^{bike}\right) * p_\theta\left(\mathbf{x}_t^{taxi} \mid \mathbf{z}_t^{taxi}\right) \\ * p_\theta\left(\mathbf{x}_t^{bus} \mid \mathbf{z}_t^{bus}\right) * p_\theta\left(\mathbf{x}_t^{v} \mid \mathbf{z}_t^{v}\right) \end{aligned} \tag{3}$$

We apply variational Bayes to learn a tractable distribution q_ϕ to approximate the true posterior p_θ, defined as follows:

$$\begin{aligned} q_\phi\left(\boldsymbol{\epsilon}_{t<T}, \mathbf{z}_{t<T} \mid \mathbf{x}_{t<T}, \mathbf{C}_{t<T}\right) = \prod_{t=1}^{T-1} q_\phi\left(\boldsymbol{\epsilon}_t \mid \mathbf{z}_{t-1}, \mathbf{x}_t, \mathbf{C}_t\right) \\ * q_\phi\left(\mathbf{z}_t \mid \boldsymbol{\epsilon}_t\right) \end{aligned} \tag{4}$$

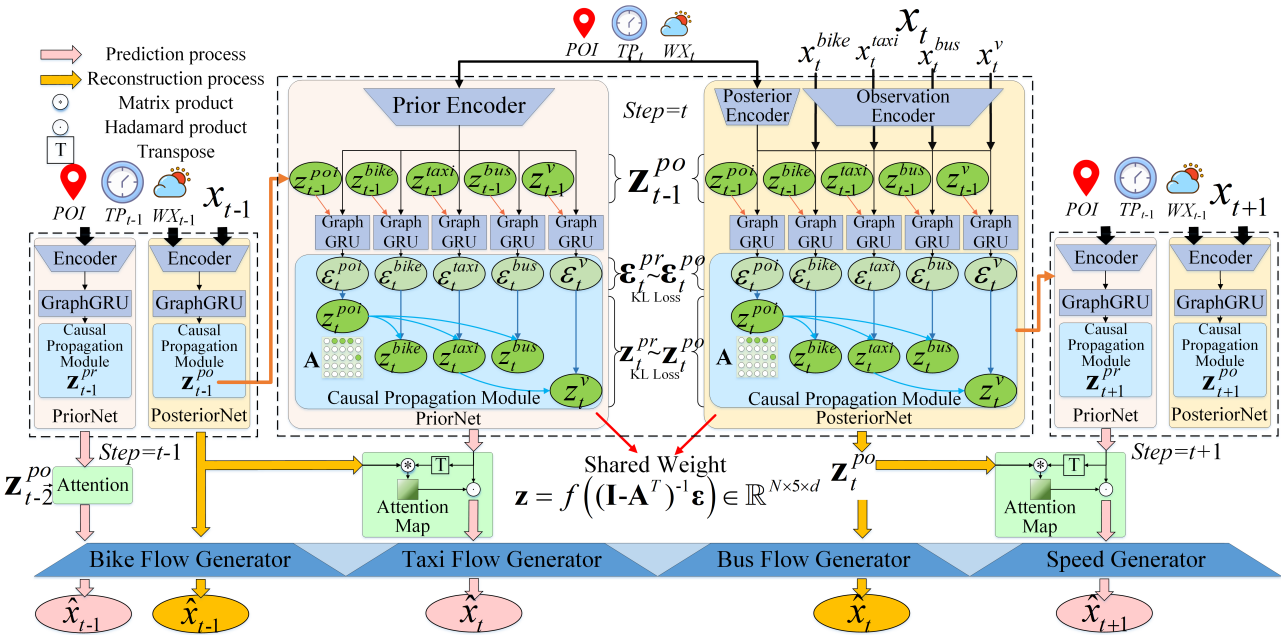

Figure 3: The architecture of CCHMM.

Causal Conditional Hidden Markov Model

To model Causal Conditional Hidden Markov Model based on the above probabilistic generative model, as shown in Fig. 3, our main tasks are as follows: (1) In the latent variable inference stage, a deep neural network is used to fit the prior distributions $p_\theta\left(\boldsymbol{\epsilon}_t, \mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{C}_t\right)$ and posterior distributions $q_\phi\left(\boldsymbol{\epsilon}_t, \mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{x}_t, \mathbf{C}_t\right)$ of latent variables to disentangle the causal representations of concepts affecting the generation of multimodal traffic observations. (2) A causal propagation module is proposed to mine the causal relationship between endogenous latent variables through a trainable causal graph, and propagate the causal effect according to the causal order. (3) In the observation data generation stage, the generator is established to approximate the conditional generation distribution $p_\theta(\mathbf{x}_t|\mathbf{z}_t)$. We utilize learnable variational distributions to approximate the true data distribution, with the aim of disentangling causal representations of physical concepts using variational inference. Compared to traditional VAE, we explicitly endow latent variables with real semantic information (i.e., causal representations of physical concepts).

Posterior Network

We use conditional information and observations to build a PosteriorNet, whose purpose is to approximate the true posterior distribution of latent variables by learning a variational posterior distribution $q_\phi\left(\boldsymbol{\epsilon}_t, \mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{x}_t, \mathbf{C}_t\right)$ using neural networks. As shown in the yellow part of Fig. 3, it consists of GraphGRU and Causal Propagation Module.

GraphGRU The progression of multimodal traffic flow has Markov property, and the evolution of latent variables is the intrinsic drivie for the spatio-temporal dependencies of multimodal traffic observations. Therefore, we use the GraphGRU to model the evolution process of system status, capturing the spatio-temporal dependencies into the exogenous latent variables. We build parameter-independent GraphGRU to learn mode-specific patterns for each traffic modes, defined as follows:

$$
\begin{aligned}
\mathbf{s}_t^{po,i} &= \mathrm{FC}(\mathbf{C}_t||\mathbf{x}_t^i) \\
\mathbf{r}_t^{po,i} &= \sigma(\mathbf{W}_r^i \star_G(\mathbf{s}_t^{po,i}||\mathbf{z}_{t-1}^{po,i}) + \mathbf{b}_r^i) \\
\mathbf{u}_t^{po,i} &= \sigma(\mathbf{W}_u^i \star_G(\mathbf{s}_t^{po,i}||\mathbf{z}_{t-1}^{po,i}) + \mathbf{b}_u^i) \\
\tilde{\mathbf{h}}_t^{po,i} &= \tanh(\mathbf{W}_h^i \star_G(\mathbf{s}_t^{po,i}||(\mathbf{r}_t^{po,i} \odot \mathbf{z}_{t-1}^{po,i})) + \mathbf{b}_h^i) \\
\boldsymbol{\epsilon}_t^{po,i} &= \mathbf{u}_t^{po,i} \odot \mathbf{z}_{t-1}^{po,i} + (1 - \mathbf{u}_t^{po,i}) \odot \tilde{\mathbf{h}}_t^{po,i}
\end{aligned}
\tag{5}
$$

where $i \in \{poi, bike, taxi, bus, v\}$ denotes physical concept of interest, $||$ denotes concatenate operation, σ denotes sigmoid funtion, $\mathbf{C}_t \in \mathbb{R}^{N \times c_c}$ is conditional information, $\mathbf{x}_t^i \in \mathbb{R}^{N \times c_i}$ is the observation of the i-th mode, c_i is the number of traffic flow channels of the i-th mode, $\mathbf{z}_{t-1}^{po,i} \in \mathbb{R}^{N \times d}$ is the posterior endogenous latent variable at $t-1$, $\boldsymbol{\epsilon}_t^{po,i} \in \mathbb{R}^{N \times d}$ is posterior exogenous latent variable at t, and $\mathbf{W}, \mathbf{b}$ are the parameters of graph convolution. The graph convolution defined by $\mathbf{W} \star_G (\mathbf{X}) + \mathbf{b} = (\mathbf{I} + \mathbf{D}^{-1/2}\mathbf{G}\mathbf{D}^{-1/2})\mathbf{X}\mathbf{W} + \mathbf{b}$, where $\mathbf{G} \in \mathbb{R}^{N \times N}$ is the distance adjacency matrix of regions, $\mathbf{D}_{ii} = \sum_j \mathbf{G}_{ij}$ and N is the number of regions. Then, we calculate the mean and log-variance of $\boldsymbol{\epsilon}_t^{po}$ by using separate fully connected layers for each traffic modes to obtain the posterior distribution of exogenous latent variables $q_\phi\left(\boldsymbol{\epsilon}_t \mid \mathbf{z}_{t-1}, \mathbf{x}_t, \mathbf{C}_t\right)$.

Causal Propagation Module Concepts that affect the generation of observations are naturally causally related. Therefore, endogenous latent variables, as semantic representations of concepts, also have causal relationships. We

propose a causal propagation module to transform independent exogenous variables into causal endogenous variables and leverage a learnable causal graph to mine their causal relationships.

The linear SCM is defined as $\mathbf{z} = \mathbf{A}^T\mathbf{z} + \boldsymbol{\epsilon} = (\mathbf{I}-\mathbf{A}^T)^{-1}\boldsymbol{\epsilon}$. We add parameter-independent nonlinear transformations for each traffic modes to improve the representation ability. In this paper, the causal propagation module is defined as:

$$\begin{aligned}
\tilde{\mathbf{A}} &= \mathrm{ReLU}\left(\tanh\left(\alpha \mathbf{W}_{\mathbf{A}}\right)\right) \in \mathbb{R}^{5\times 5} \\
\mathbf{h}_t^{po} &= \left(\mathbf{I} - \tilde{\mathbf{A}}^T\right)^{-1} \boldsymbol{\epsilon}_t^{po} \in \mathbb{R}^{N\times 5\times d} \\
\mathbf{z}_t^{po,i} &= f_i\left(\left[\mathbf{h}_t^{po}\right]_{[:,i,:]}\right) = \mathrm{FC}_i\left(\tanh\left(\mathrm{FC}_i\left(\left[\mathbf{h}_t^{po}\right]_{[:,i,:]}\right)\right)\right)
\end{aligned} \tag{6}$$

where $\mathbf{W}_{\mathbf{A}} \in \mathbb{R}^{5\times 5}$ denotes a learnable parameter, α is a hyper-parameter for controlling the saturation rate of the activation function. ReLU regularizes the parameter matrix to ensure sparsity and non-negativity. $\tilde{\mathbf{A}}$ is the causal graph of endogenous latent variables, where $\tilde{\mathbf{A}}_{ij}$ represents the causal effect of the parent variable $\mathbf{z}_i$ on the child variable $\mathbf{z}_j$. Therefore, when the graph nodes are sorted in topological order, the matrix $\tilde{\mathbf{A}}$ is strictly upper triangular. Then, we calculate the mean and log-variance by using separate fully connected layers for each traffic modes.

Prior Network

Previous unsupervised disentangled representation learning based on VAE regularizes the posterior of the latent variables with a standard Multivariate Gaussian prior, which greatly limits the expression ability of the model. Unsupervised disentangled representation learning can not guarantee the model identifiability due to the lack of inductive bias (Locatello et al. 2019). To improve the identifiability of the model, we build a PriorNet based on conditional information, which aims to model the physical rules of the concepts of interest that naturally exist in the system, and use a learnable prior distribution to approximate this rules. As shown in the pink part of Fig. 3, the PirorNet is similar in structure to the PosteriorNet, which is composed of GraphGRU and causal propagation module.

GraphGRU The PriorNet only inputs the conditional information of the current system, calculates prior exogenous latent variables $\boldsymbol{\epsilon}_t^{pr}$ according to Eq.5, and then obtains the prior distribution of exogenous latent variables $p_\theta\left(\boldsymbol{\epsilon}_t \mid \mathbf{z}_{t-1}, \mathbf{C}_t\right)$ by calculating the mean and log-variance.

Causal Propagation Module The PriorNet and the PosteriorNet share a causal propagation module. We argue that causality is a stable natural phenomenon that does not change with time or space, thus globally sharing a causal graph and nonlinear transformation. We calculate the prior endogenous latent variables $\mathbf{z}_t^{pr}$ according to Eq.6, and then obtain the prior distribution of the endogenous latent variables $p_\theta\left(\mathbf{z}_t \mid \boldsymbol{\epsilon}_t\right)$ by calculating the mean and log-variance.

Generator

We build the generator using two fully connected layers to parameterize the conditional distribution of the generative models $p_\theta\left(\mathbf{x}_t \mid \mathbf{z}_t\right)$ defined in Eq.3. As shown in Figure 3, a generator is globally shared. The generator is shared globally as shown in Fig. 3. The results of generative models have different meanings depending on the type of $\mathbf{z}$.

Reconstruction As shown by the yellow arrow in Fig. 3, the PosteriorNet takes the current observations as part of the input. So when generating data using the posterior endogenous latent variables $\mathbf{z}_t^{po}$, the output is the reconstruction result, represented as $\hat{\mathbf{x}}_t^{i,rec} = \mathrm{Generator}_i(\mathbf{z}_t^{po,i})$, $i \in \{bike, taxi, bus, v\}$.

Prediction The PriorNet only utilizes the current conditional information to fit the prior distribution and does not involve current observations. Therefore, when generating data using the prior endogenous latent variables $\mathbf{z}_t^{pr}$, the output is the prediction result. Based on the Markov property of sequence generation, we leverage a simple attention mechanism to weight the current prior endogenous latent variables with the previous posterior, which can further improve the effect of prediction. The attention mechanism is defined as:

$$\tilde{\mathbf{z}}_t^{pr} = \mathrm{softmax}\left(\mathbf{z}_{t-1}^{po}\mathbf{W}_{att}(\mathbf{z}_t^{pr})^T\right)\mathbf{z}_t^{pr} \in \mathbb{R}^{N\times 5\times d} \tag{7}$$

where $\mathbf{W}_{att} \in \mathbb{R}^{d\times d}$ is the learnable parameter. Then $\tilde{\mathbf{z}}_t^{pr}$ is fed into the generator to obtain the prediction result, represented as $\hat{\mathbf{x}}_t^{i,pred} = \mathrm{Generator}_i(\tilde{\mathbf{z}}_t^{pr,i})$, $i \in \{bike, taxi, bus, v\}$.

Learning Strategy

We propose a mutually supervised training method for the PriorNet and PosteriorNet, which benefits the model to approximate the physical rules of concepts of interest, while helping the to identifiably disentangle causal representations. Based on variational inference, we use a neural network to learn a tractable distribution q_ϕ to approximate the true posterior distribution p_θ. Given a dataset $\mathcal{D}$, the Evidence Lower Bound (ELBO) of CCHMM is as follows:

$$\begin{aligned}
\mathcal{L}_{ELBO} &= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{q_\phi}\left[\log\left(\frac{p_\theta(\mathbf{x}_{t<T}, \boldsymbol{\epsilon}_{t<T}, \mathbf{z}_{t<T}|\mathbf{C}_{t<T})}{q_\phi(\boldsymbol{\epsilon}_{t<T}, \mathbf{z}_{t<T}|\mathbf{x}_{t<T}, \mathbf{C}_{t<T})}\right)\right]\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\sum_{t=1}^{T-1}\mathcal{L}_t^{q_\phi, p_\theta}\right] \\
\mathcal{L}_t^{q_\phi, p_\theta} &= \mathbb{E}_{q_\phi\left(\boldsymbol{\epsilon}_t, \mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{x}_t, \mathbf{C}_t\right)}\left[\log\left(p_\theta(\mathbf{x}_t|\mathbf{z}_t)\right)\right] \\
&\quad - D_{KL}\left[q_\phi\left(\boldsymbol{\epsilon}_t, \mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{x}_t, \mathbf{C}_t\right)||p_\theta\left(\boldsymbol{\epsilon}_t, \mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{C}_t\right)\right]
\end{aligned} \tag{8}$$

We rewrite Eq. 6 as $\mathbf{z}_t = \varphi_{\mathbf{w}}(\epsilon_t)$, where $\mathbf{w}$ is the parameter of the causal propagation module and $\varphi_{\mathbf{w}}$ is invertible. Therefore, we reformulate the of the prior and posterior distributions with the Dirac delta function $\delta(\cdot)$, represented as follows:

$$\begin{aligned}
q_\phi\left(\boldsymbol{\epsilon}_t, \mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{x}_t, \mathbf{C}_t\right) &= q_\phi\left(\boldsymbol{\epsilon}_t|\mathbf{z}_{t-1}, \mathbf{x}_t, \mathbf{C}_t\right)\delta\left(\mathbf{z}_t = \varphi_{\mathbf{w}}(\boldsymbol{\epsilon}_t)\right) \\
&= q_\phi\left(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{x}_t, \mathbf{C}_t\right)\delta\left(\boldsymbol{\epsilon}_t = \varphi_{\mathbf{w}}^{-1}(\mathbf{z}_t)\right) \\
p_\theta\left(\boldsymbol{\epsilon}_t, \mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{C}_t\right) &= p_\theta\left(\boldsymbol{\epsilon}_t|\mathbf{z}_{t-1}, \mathbf{C}_t\right)\delta\left(\mathbf{z}_t = \varphi_{\mathbf{w}}(\boldsymbol{\epsilon}_t)\right) \\
&= p_\theta\left(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{C}_t\right)\delta\left(\boldsymbol{\epsilon}_t = \varphi_{\mathbf{w}}^{-1}(\mathbf{z}_t)\right)
\end{aligned} \tag{9}$$

Models	Bike			Taxi			Bus			Speed		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
HMM	6.323	13.102	24.812%	5.014	8.614	26.885%	6.797	13.033	21.204%	1.376	2.226	4.313%
CCRNN	5.353	11.341	21.112%	4.758	8.710	24.739%	6.671	13.452	20.263%	1.556	2.560	4.891%
DMSTGCN	5.267	9.975	21.504%	4.587	7.949	24.204%	6.361	12.110	19.957%	1.407	2.251	4.359%
AGCRN	5.018	9.357	20.381%	4.561	7.899	23.988%	6.558	12.508	19.986%	1.367	2.158	4.276%
DGCRN	4.937	9.143	20.328%	4.536	7.898	23.984%	6.428	12.223	19.649%	1.415	2.287	4.409%
CCHMM(our)	**4.641**	**8.521**	**19.428%**	**4.415**	**7.626**	**23.566%**	**6.245**	**11.857**	**19.203%**	**1.243**	**1.994**	**3.858%**

Table 1: Performance comparison with other models.

We substitute the prior and the posterior distributions in Eq. 9 and reformulate $\mathcal{L}_t^{q_\phi,p_\theta}$ as:

$$\begin{aligned}\mathcal{L}_t^{q_\phi,p_\theta} =& \mathbb{E}_{q_\phi(\mathbf{z}_t|\mathbf{z}_{t-1},\mathbf{x}_t,\mathbf{C}_t)}\left[\log\left(p_\theta(\mathbf{x}_t|\mathbf{z}_t)\right)\right] \\ &- D_{KL}\left[q_\phi\left(\boldsymbol{\epsilon}_t|\mathbf{z}_{t-1},\mathbf{x}_t,\mathbf{C}_t\right)||p_\theta\left(\boldsymbol{\epsilon}_t|\mathbf{z}_{t-1},\mathbf{C}_t\right)\right] \\ &- D_{KL}\left[q_\phi\left(\mathbf{z}_t|\mathbf{z}_{t-1},\mathbf{x}_t,\mathbf{C}_t\right)||p_\theta\left(\mathbf{z}_t|\mathbf{z}_{t-1},\mathbf{C}_t\right)\right]\end{aligned} \tag{10}$$

where, the first term is the reconstruction loss, and the last two terms are the KL divergence of the exogenous and endogenous latent variables, respectively.

Since the causal graph has the property of being acyclic, it is necessary to increase the acyclic constraint (Yu et al. 2019) of $\tilde{\mathbf{A}}$, expressed as $h(\tilde{\mathbf{A}}) = \mathrm{tr}\left[(I + \tilde{\mathbf{A}} \circ \tilde{\mathbf{A}})^n\right] - n$. In addition, we use L2-norm as the predicted loss, defined as $\mathcal{L}^{pred} = \mathbb{E}_{\mathcal{D}}\left[\sum_{t=1}^{T-1}\left\|\hat{\mathbf{x}}_t^{pred} - \mathbf{x}_t\right\|_2^2\right]$. In summary, the total loss function of CCHMM is defined as follows:

$$\mathcal{L} = -\mathcal{L}_{ELBO} + \mathcal{L}_{pred} + \lambda h(\tilde{\mathbf{A}}) \tag{11}$$

where λ is hyper-parameter for controlling the loss balance.

Experiments

We evaluate the performance of our model on real world traffic datasets and compare with some recent compelling methods for traffic flow prediction[1]. Further, a comprehensive ablation study shows the effectiveness of each component of our model.

Dataset

XC-Trans:The XC-Trans dataset contains order records of three traffic modes(bike, bus and taxi) from June 1st 2021 to December 31th 2021 in Xicheng District, Beijing. The researched region is split into 175 non-overlapping subregions. We statistics the inflow and outflow for each traffic modes in all of the subregions.

XC-Speed:The XC-speed dataset contains speed records of main roads from June 1st 2021 to December 31th 2021 in Xicheng District, Beijing. We use the average speed of road segments within each region to represent the regional speed.

Besides, corresponding meteorological information, time position and POI data are collected as conditional information. We split this dataset with a 30-minute interval to obtain 11753 samples. we use three-hour historical data to predict the next 30-minute data. 60% of the data is used for training, 20% is used for validating and the rest is used for testing.

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

Experimental Settings

We compare our framework with the following methods. 1)**HMM**(Li et al. 2021c): It uses multimodal information to achieve robust prediction of irreversible disease at an early stage. 2) **AGCRN**(Bai et al. 2020): It employs an adaptive graph and integrates GRU with graph convolutions. 3) **CCRNN**(Ye et al. 2021): It employs coupled layer-wise graph convolution layer to capture the multi-level spatial dependence and temporal dynamics simultaneously. 4) **DGCRN**(Li et al. 2021b): It generates a dynamic graph by combining the predefined adjacency matrix and input features. 5) **DMSTGCN**(Han et al. 2021): It designs an adaptive graph construction method to learn the time-specific spatial dependencies of road segments.

Overall Comparison

We evaluate the performance of methods with Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error(MAPE). Table 1 presents the overall prediction performances which are the averaged results over three independent experiments. There is no method compatible with all traffic modes except us.

The baseline models focus on adaptively or dynamically generating graph structures, while our model pay more attention to modeling the causality between latent semantic variables in traffic system. Due to the lack of modeling spatial dependency and causality, the HMM model shows the worst performance. The model based on dynamic graph(e.g. DGCRN) perform better than models based on adaptive graph(e.g. AGCRN). Besides, it can be observed that our model outperforms baseline models consistently and overwhelmingly. Especially in speed prediction, our CCHMM brings about 10% improvements to the best results in all metrics due to the causality of speed factor is more clear

Ablation Study

To evaluate the effectiveness of key components, we conduct comprehensive ablation experiments. For PriorNet, we design four variants: 1)w/o GRU: This variant replaces GraphGRU with GCN. The prior of latent variables is only generated from conditional information, which means discarding long-term temporal dependencies. 2)w/o GCN: This variant removes GCN in GraphGRU, which means discarding spatial dependencies. 3) w/o Cond: This variant removes conditional information. Note that we consider ϵ as exogenous variables which are relevant to conditional information. Removing conditional information is equivalent to removing

Models	Bike			Taxi			Bus			Speed		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
w/o GRU	10.41	22.79	39.89%	9.40	17.62	47.02%	10.74	21.50	28.97%	1.99	3.06	6.40%
w/o GCN	6.31	12.67	24.86%	5.27	9.38	27.11%	6.75	12.93	20.94%	1.43	2.33	4.51%
w/o Cond	5.76	11.15	22.96%	5.54	9.93	28.46%	7.11	13.88	21.55%	1.47	2.39	4.61%
w/o Prior	5.49	10.64	21.83%	5.07	9.05	26.19%	6.88	13.26	21.10%	1.47	2.43	4.57%
Entangle	5.53	10.43	22.54%	5.15	9.05	26.90%	6.96	13.19	21.77%	1.55	2.52	4.86%
w/o SCM	5.19	9.50	21.66%	4.90	8.47	26.33%	6.63	12.60	20.47%	1.44	2.35	4.48%
w/o Nonlinear	4.95	8.70	20.90%	4.53	7.76	24.25%	6.52	12.43	20.11%	1.34	2.17	4.19%
CCHMM	**4.64**	**8.52**	**19.42%**	**4.41**	**7.62**	**23.56%**	**6.24**	**11.85**	**19.20%**	**1.24**	**1.99**	**3.85%**

Table 2: Results of ablation study.

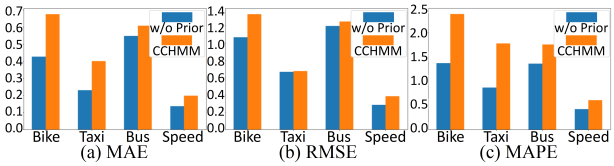

Figure 4: Comparison of Reconstruction performance of w/o Prior and CCHMM.

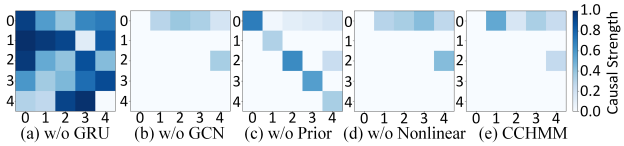

Figure 5: Comparison of Reconstruction performance of w/o Prior and CCHMM.

PriorNet and generating latent variables directly from observation data in PostierNet. 4) w/o Prior: This variant removes the PriorNet but retains conditional information. Different from variant 3, the latent variables are generated from both conditional information and observation with SCM. For the causal propagation module, we design three variants: 5) Entangle: There is only one latent variable in this variant. 6) w/o SCM: This variant removes SCM, which means the latent variables are directly generated from conditional information and observation. 7) w/o Nonlinear: This variant replaces the non-linear transformation with linear transformation in SCM. Note that except for variant 3 and variant 4 that use additional FC layers for prediction, other networks use generator to obtain prediction results.

The performance of ablation experiments is shown in Table 2. We can find that variant 1 and 2 perform worst of all due to the lack of spatial and temporal dependencies. The performance of variant 3 shows the necessity of conditional information. In fact, the exogenous variables only affect the system, but are not constrained by the system. It means that we can only determine them by conditional information. Eventually, the model without conditional information degenerates into an ordinary sequence disentangled representation learning model. In variant 4, we drop the PriorNet. The role of the PriorNet is to obtain the stable rules of physical concepts, while the PosteriorNet is designed for obtaining disentangled representations from observation data and conditional information. Posterior collapse may occur in the absence of prior supervision, resulting in failure to obtain a stable and effective causal representation. An evidence is shown in Fig. 4, the reconstruction loss of variant 4 is generally lower than our CCHMM, which means that the model prefers to learn a representation for reconstruction rather than disentangling.

For causal propagation module, the model with disentangle latent variables perform better than the entangle one, which means that VAE-based structures decouple the latent variables to some extent. Since there is no restriction of causal structure, it suffers from spurious correlation. The most obvious consequence is that the speed prediction performance is reduced by 15%. Besides, the performance of variant 7 linear model is insufficient to express causal relationships in complex scenarios.

In addition, for each model with causal propagation module, we initialize the causal graph as an upper triangular matrix subject to standard normal distribution. As shown in Fig. 5, it can be observed that the variant 1 failed to learn a stable causal relationship. The model without GCN and the one without non-linear transformation learnt a causal graph similar to our CCHMM. Particularly, the model without PriorNet learnt an causal graph with large diagonal elements. It means that the model failed to learn representations of physical concepts that conform to causality.

Conclusion

In this paper, we analyze the core physical concepts affecting the generation of multimodal traffic flow and disentangle the concepts of interest into three groups: regional attraction factor, the transportation demand factor and traffic speed factor. We infer causal representations of these concepts from conditional information and observations of the current system based on variational inference and structural causal model, and mine their causal relationships by using learnable causal graphs. For the data generation stage, we feed the prior causal representation into the generator to generate predictions. Extensive experiments show that all metrics of CCHMM are optimal, which reveal that it is crucial to introduce causal theory into spatio-temporal sequence analysis. In future work we further explore causal discovery and refine causal relationships in multimodal traffic systems.

References

Bai, L.; Yao, L.; Li, C.; Wang, X.; and Wang, C. 2020. Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in Neural Information Processing Systems*, 33: 17804–17815.

Cao, D.; Zeng, K.; Wang, J.; Sharma, P. K.; Ma, X.; Liu, Y.; and Zhou, S. 2021. BERT-Based Deep Spatial-Temporal Network for Taxi Demand Prediction. *IEEE Transactions on Intelligent Transportation Systems*.

Deng, J.; Chen, X.; Fan, Z.; Jiang, R.; Song, X.; and Tsang, I. W. 2021. The Pulse of Urban Transport: Exploring the Co-evolving Pattern for Spatio-temporal Forecasting. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(6): 1–25.

Deng, X.; and Zhang, Z. 2021. Comprehensive Knowledge Distillation with Causal Intervention. *Advances in Neural Information Processing Systems*, 34.

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.

Khemakhem, I.; Kingma, D.; Monti, R.; and Hyvarinen, A. 2020. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, 2207–2217. PMLR.

Li, C.; Bai, L.; Liu, W.; Yao, L.; and Waller, S. T. 2021a. A multi-task memory network with knowledge adaptation for multimodal demand forecasting. *Transportation Research Part C: Emerging Technologies*, 131: 103352.

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

Li, J.; Wu, B.; Sun, X.; and Wang, Y. 2021c. Causal Hidden Markov Model for Time Series Disease Forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12105–12114.

Liang, Y.; Huang, G.; and Zhao, Z. 2021. Joint Demand Prediction for Multimodal Systems: A Multi-task Multi-relational Spatiotemporal Graph Neural Network Approach. *arXiv preprint arXiv:2112.08078*.

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.

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

Liu, H.; Wu, Q.; Zhuang, F.; Lu, X.; Dou, D.; and Xiong, H. 2021b. Community-Aware Multi-Task Transportation Demand Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 320–327.

Locatello, F.; Bauer, S.; Lucic, M.; Raetsch, G.; Gelly, S.; Schölkopf, B.; and Bachem, O. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, 4114–4124. PMLR.

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.

Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N. R.; Kalchbrenner, N.; Goyal, A.; and Bengio, Y. 2021. Towards causal representation learning. *arXiv preprint arXiv:2102.11107*.

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

Wang, Q.; Guo, B.; Ouyang, Y.; Cheng, L.; Wang, L.; Yu, Z.; and Liu, H. 2021. Learning Shared Mobility-aware Knowledge for Multiple Urban Travel Demands. *IEEE Internet of Things Journal*.

Wu, Z.; Pan, S.; Long, G.; Jiang, J.; Chang, X.; and Zhang, C. 2020. Connecting the Dots: Multivariate Time Series Forecasting with Graph Neural Networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

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.; Tong, X.; and Xiong, H. 2019. Co-prediction of multiple transportation demands based on deep spatio-temporal neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 305–313.

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.

Yu, Y.; Chen, J.; Gao, T.; and Yu, M. 2019. DAG-GNN: DAG structure learning with graph neural networks. In *International Conference on Machine Learning*, 7154–7163. PMLR.

Zhou, Q.; Gu, J.; Lu, X.; Zhuang, F.; Zhao, Y.; Wang, Q.; and Zhang, X. 2021. Modeling Heterogeneous Relations across Multiple Modes for Potential Crowd Flow Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 4723–4731.