CityCAN: Causal Attention Network for Citywide Spatio-Temporal Forecasting

Chengxin Wang
cwang@comp.nus.edu.sg
National University of Singapore
Singapore

Yuxuan Liang
yuxliang@outlook.com
The Hong Kong University of Science and Technology (Guangzhou)
China

Gary Tan
gtan@comp.nus.edu.sg
National University of Singapore
Singapore

ABSTRACT

Citywide spatio-temporal (ST) forecasting is a fundamental task for many urban applications, including traffic accident prediction, taxi demand planning, and crowd flow forecasting. The goal of this task is to generate accurate predictions concurrently for all regions within a city. Prior works take great effort on modeling the ST correlations. However, they often overlook intrinsic correlations and inherent data distribution across the city, both of which are influenced by urban zoning and functionality, resulting in inferior performance on citywide ST forecasting. In this paper, we introduce CityCAN, a novel causal attention network, to collectively generate predictions for every region of a city. We first present a causal framework to identify useful correlations among regions, filtering out useless ones, via an intervention strategy. In the framework, a Global Local-Attention Encoder, which leverages attention mechanisms, is designed to jointly learn both local and global ST correlations among correlated regions. Then, we design a citywide loss to constrain the prediction distribution by incorporating the citywide distribution. Extensive experiments on three real-world applications demonstrate the effectiveness of CityCAN.

CCS CONCEPTS

• **Information systems** → **Spatial-temporal systems**; • **Applied computing** → **Forecasting**.

KEYWORDS

Spatio-temporal Data Mining, Causal Intervention, Attention

ACM Reference Format:
Chengxin Wang, Yuxuan Liang, and Gary Tan. 2024. CityCAN: Causal Attention Network for Citywide Spatio-Temporal Forecasting. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM '24), March 4–8, 2024, Merida, Mexico.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3616855.3635764

1 INTRODUCTION

To build Intelligent Transportation Systems (ITS), numerous sensors are widely placed in cities to capture traffic conditions [21], producing massive *spatio-temporal (ST) data*, as depicted in Fig. 1 (a). Foreseeing citywide ST data, such as traffic accidents, crowd flows, and crowd densities, is crucial for ITS development. It can facilitate various urban applications, such as assisting transportation managers in mitigating accidents [3, 11, 53], guiding car-sharing companies in vehicle allocation [12, 18, 61], and aiding drivers in selecting optimal routes [8, 17, 19].

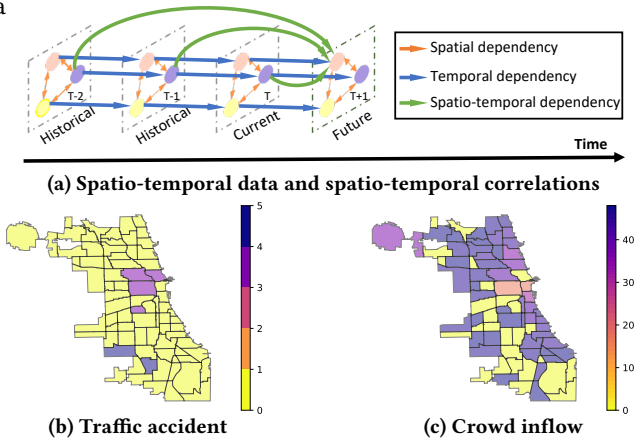


Figure 1: (a) A illustration of the spatio-temporal data and three types of dependency. (b-c) A visualization of citywide traffic accidents and crowd inflow in Chicago on July 1, 2021 at 14:00. Traffic conditions are influenced by urban zoning.

One key characteristic of citywide ST data is *spatio-temporal (ST) correlations*, illustrated in Fig. 1 (a). Specifically, a target region's conditions are influenced by three dependencies: spatial (represented by the orange line), temporal (blue line), and spatio-temporal (ST) (green line). In the era of big data, researchers have proposed many data-driven methods, especially deep learning approaches, to capture these ST correlations. Most works [47, 60] address spatial and temporal dependency separately, neglecting the direct ST dependency. They typically capture spatial dependencies via convolutional neural networks (CNNs) [5, 75] or graph neural networks (GNNs) [14, 58], and exploit temporal dependencies with recurrent mechanisms (RNNs) [42, 65] or attention mechanisms [68, 69, 73]. To fully exploit the ST correlations, more recent approaches model the spatial and temporal dependency simultaneously via local ST graphs [43], pyramid CNNs [29] or ST enhanced mechanisms [48]. Despite recent advances in ST modeling, two major challenges persist in forecasting citywide ST data:

Challenge I: How to identify the useful correlations among regions across time? Studies have shown that urban zoning and functionality influence the citywide ST correlations [7, 30]. To incorporate urban functionality into citywide forecasting, previous studies [28, 30] integrate geographical features (e.g., Points of Interest) as auxiliary inputs to ST networks. However, these methods reply on ST

WSDM '24, March 4–8, 2024, Merida, Mexico

ACM ISBN 979-8-4007-0371-3/24/03
https://doi.org/10.1145/3616855.3635764

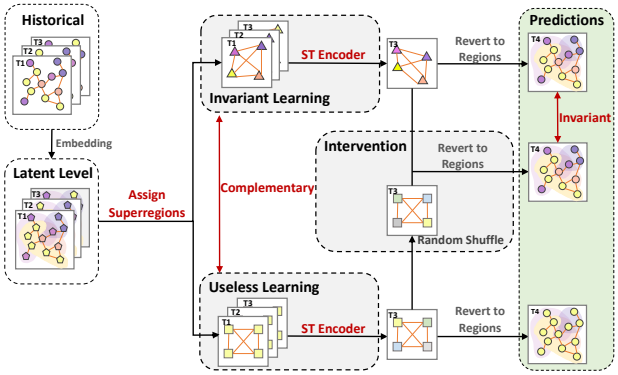


Figure 2: Insight of the network.

networks to learn spatial correlations, often leading to an over-generalized consideration of ST correlations across all regions. In reality, one region's future conditions is largely influenced by regions with useful correlations, rather than every region in the city. For example, the work area traffic is intrinsically correlated to residential areas due to daily commutes but shows limited correlation with agricultural zones. Thus, instead of broadly capturing all ST correlations, we emphasize identifying and utilizing useful correlations among regions to enhance citywide ST forecasting.

Challenge II: How to constrain the predictions to align with actual citywide distributions? The characteristics of urban zoning and functionality give citywide ST data a distinctive distribution [40, 64], as illustrated in Fig 1 (b-c). For instance, most traffic events (e.g., accidents) take place in urban areas and they rarely occur in rural areas [3, 47]; taxi flows are concentrated in downtown districts and sparse in other boroughs [16, 77]. Previous works [5, 67] prioritize regions with large event numbers, often employing region-wise losses, such as MSE, RMSE, to optimize neural networks. Some of them [41, 47] even specifically amplify the impact of high-event regions through re-weighted loss strategies. However, these works force networks to calculate errors that are skewed towards the regions with large event numbers, overlooking these with fewer events. Thus, they may cause the network to generate predictions that considerably deviate from actual citywide distributions.

In this paper, we propose a **C**ausal **A**ttention **N**etwork (CityCAN) for citywide ST forecasting. Given useful correlations are influenced by urban zoning, we argue these correlations are invariant spatial correlations among regions over time. Thus, *to address challenge I*, CityCAN employs a causal framework (depicted in Fig. 2) to learn the useful correlations, while ignoring its complementary correlations (i.e., useless correlations). In CityCAN, regions with invariant/useless correlations are assigned to useful/useless superregions for invariant/useless learning branches with two complementary superregion matrices. Then, the useful correlations can be identified by pushing the predictions from the invariant learning branch and intervention branch to be invariant, regardless of changes in useless representations learned from the useless learning branch. Enhancing the ST modeling within these branches, we propose a novel Global Local-Attention Encoder (GLAE) as the ST Encoder to jointly encode spatial and temporal dependencies via local and global ST attentions. *To tackle challenge II*, we design a citywide loss that penalizes the network from a global perspective, i.e., on the city level. Specifically, it constrains the predictions on spatial dimension aligns closely with the true spatial distribution by considering all regions in the city collectively. In other words, it measures the distribution similarity between predictions and future conditions across all regions. Overall, we summarize our contributions as follows:

- We propose CityCAN, a causal attention network for citywide ST forecasting, which leverages causal theory to uncover useful spatial correlations over time.
- We introduce a Global Local-Attention Encoder (GLAE) for better spatio-temporal correlation modeling.
- We design a citywide loss, which constrains the prediction distribution, leading to improved citywide ST forecasting.
- Experiments show CityCAN outperforms state-of-the-art methods on four datasets in three practical applications.

2 PRELIMINARIES

Definition 1 (Region): The area of interest, i.e., city, is divided into N regions based on their longitude and latitude [47]. These regions can be either regular or irregular in shape.

Definition 2 (Traffic Condition & Traffic Features): Traffic conditions are traffic-related conditions, such as the risk level for traffic accident data, inflow/outflow for crowd flow data and the count for crowd density data. The features of these traffic conditions are traffic features X. Given a time interval t, $X_t \in \mathbb{R}^{N\times d}$, where d is the dimension of the traffic features.

Problem Statement Given observed traffic features with T time intervals $I_{1:T} = (I_1, I_2, \ldots, I_T) \in \mathbb{R}^{N\times T\times d_i}$, spatial adjacency matrix of regions $\mathbf{A} \in \mathbb{R}^{N\times N}$, the goal is to generate interested traffic features for all regions for next T_{pred} time intervals, i.e., $\hat{O}_{T+1:T+T_{pred}} = (O_{T+1}, O_{T+2} \ldots, O_{T+T_{pred}}) \in \mathbb{R}^{N\times T_{pred}\times d_o}$. Note that the observed traffic features may have more information than the interested traffic features, i.e., $d_i \geq d_o$.

3 CITYCAN

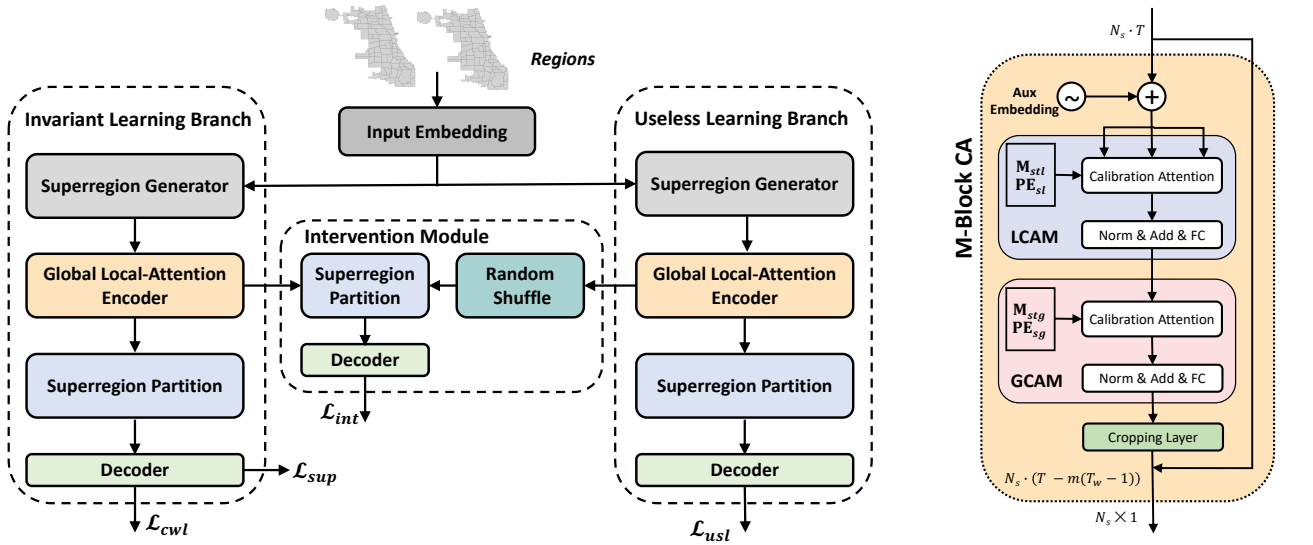
In this section, we present our CityCAN, as shown in Fig. 3, which employs two strategies to tackle citywide ST forecasting: a causal framework to identify useful ST correlations (Section 3.1) and a citywide loss to constrain the prediction distribution (Section 3.2).

3.1 Causal Learning for Citywide Forecasting

Due to the urban zoning and functionality, despite ST correlations in citywide data can be dynamic in a short period (e.g., days), invariant spatial correlations among city regions (e.g., correlations between residential and school areas) do exist over time. We treat these invariant correlations as useful correlations. To identify these correlations, inspired by classification tasks [39, 44] that adopt causal theory to disentangle the relevant and irrelevant features, we take a causal look at the citywide ST forecasting and propose a causal learning strategy for this regression task.

3.1.1 ***Causal Learning Strategy.*** To identify the useful (i.e., invariant) correlations, CityCAN, as shown in Fig. 3 (a), employs a causal framework, which contains an Input Embedding, an Invariant Learning Branch (ILB), a Useless Learning Branch (ULB), and an Intervention Module (IM).

Input Embedding transforms raw inputs into the latent level for superregions assignment. We first embed the traffic features $X_{1:T}$



(a) CityCAN (b) Global Local-Attention Encoder

Figure 3: (a) Overview of CityCAN, where invariant and useless correlations among regions are disentangled via the causal learning strategy. (b) Global Local-Attention Encoder (GLAE) learns useful and useless ST features based on invariant and useless correlations. It has the same architecture but different parameters in the invariant and useless learning branches.

into high-level representations through a 2D convolution with kernel size $(1, 1)$. To inject the space-time location for each region, we extend the positional embedding [71] to ST positional embeddings. Specifically, for a region n at time t (denoted as $B(n, t)$), we define its absolute space-time position as $(n + tN)$, where the index of n and t starts from 0. By adding the representations and ST positional embeddings, we obtain final features $\mathcal{H} \in \mathbb{R}^{N \times T \times d_h}$ for ILB and ULB to assign useful and useless superregions, where d_h is the feature dimension.

Invariant & Useless Learning Branch (ILB & ULB) work on learning the invariant and useless correlations among regions, respectively. They share the same architecture, which includes a Superregion Generator, a Global Local-Attention Encoder, a Superregion Partition, and a Decoder.

Superregion Generator groups regions with correlations between each other into one superregion. To identity useful correlations and filtering out useless ones, we introduce two learnable superregion matrices, i.e., useful superregion matrix $\mathbf{G}_c$ and useless superregion matrix $\mathbf{G}_u$. They are derived from the correlations observed among regions in the training data. To group correlated/uncorrelated regions to the useful/useless superregions, we apply the matrices to the original regions and their corresponding adjacent relationships:

$$\tilde{\mathcal{H}}_c = \mathcal{H}\mathbf{G}_c \quad \tilde{A}_c = \mathbf{G}_c^{\mathsf{T}} A \mathbf{G}_c; \quad \tilde{\mathcal{H}}_u = \mathcal{H}\mathbf{G}_u \quad \tilde{A}_u = \mathbf{G}_u^{\mathsf{T}} A \mathbf{G}_u \quad (1)$$

where $\tilde{\mathcal{H}}_c, \tilde{\mathcal{H}}_u \in \mathbb{R}^{N_s \times T \times d_h}$, $\mathbf{G}_c, \mathbf{G}_u \in \mathbb{R}^{N \times N_s}$, $\tilde{A}_c, \tilde{A}_u \in \mathbb{R}^{N_s \times N_s}$ is the learned adjacent metrics of superregions, $N_s = ceil(N/r)$ is total number of superregions in space dimension, r denotes the *region reduction parameter*, and T is the transposition operation. We ensure the useful and useless correlations are complementary to each other, and thus let superregion metrics satisfy $\mathbf{G}_c + \mathbf{G}_u = \mathbf{1}$, where $\mathbf{1}$ is the all-one matrix.

Thus, we can obtain $L = N_s \times T$ useful/useless superregions with their corresponding useful/useless features $\tilde{\mathcal{H}}_c/\tilde{\mathcal{H}}_u$ and adjacent relationships $\tilde{A}_c/\tilde{A}_u$ in ILB/ULB. In ILB, the GLAE (our ST Encoder) takes $\tilde{\mathcal{H}}_c$ and $\tilde{A}_c$ as inputs, while in the ULB, it uses $\tilde{\mathcal{H}}_u$ and $\tilde{A}_u$ as inputs. GLAE works on capturing ST correlations, either within the useful superregions in ILB or the useless ones in ULB. It produces ST representations $\mathbf{h}_c$ for useful superregions and $\mathbf{h}_u$ for the useless ones (More details in Section 3.1.2). These representations can be easily mapped back to their original regions via a *Superregion Partition* using the superregion matrices $\mathbf{G}_c, \mathbf{G}_u$:

$$\vec{\mathcal{H}}_c = \mathbf{h}_c \mathbf{G}_c^{\mathsf{T}} \qquad \vec{\mathcal{H}}_u = \mathbf{h}_u \mathbf{G}_u^{\mathsf{T}} \quad (2)$$

where $\vec{\mathcal{H}}_c, \vec{\mathcal{H}}_u \in \mathbf{R}^{N \times d_h}$. Then, given useful ST representation $\vec{\mathcal{H}}_c$ and useless ST representation $\vec{\mathcal{H}}_u$ of original regions, we use the fully-connected layers as the *Decoder* in ILB and ULB to generate the predictions and useless predictions $\hat{O}$, $\hat{O}_{usl} \in \mathbb{R}^{N \times T_{pred} \times d_o}$, where d_o is a task-specific dimension of traffic features. Note that since GLAE is an attention-based encoder, it reduces r^2 complexity given the region reduction parameter r.

Intervention Module (IM) aims to eliminate the influence of useless representations by providing implicit interventions on the latent level. Inspired by [44], we first generate interventions using a *Random Shuffle* operation, which randomly collects useless representations from different useless superregions. These random interventions are then concatenated with the useful representation $\mathbf{h}_c$ to generate the intervened predictions $\hat{O}_{int}$ via the Superregion Partition with $\mathbf{G}_c$ and the Decoder. Then, we encourage the invariance between the intervened predictions $\hat{O}_{int}$ and the predictions $\hat{O}$ obtained from the ILB to mitigate the impact of useless features through an intervention loss $\mathcal{L}_{\text{int}}$:

$$\mathcal{L}_{\text{int}}\left(O, \hat{O}_{int}\right) = \frac{1}{N} \sum_{n=0}^{N} \left(O_n - \Phi_{spd}\left(\mathbf{h}_c \,||\, f_{rs}(\mathbf{h}_u)\right)\right)^2 \quad (3)$$

where f_{rs} denotes the random shuffle operation, Φ_{spd} represents operations in Superregion Partition and Decoder, and $||$ refers to the concatenation function. To this end, CityCAN can fully exploit the useful correlations by ignoring the influence of useless correlations.

Losses for Causal Learning Except for the intervention loss $\mathcal{L}_{\text{int}}$, we also introduce supervised loss and useless loss to disentangle the useful features and useless features for boundless traffic condition values. Supervised loss $\mathcal{L}_{\text{sup}}$ estimates predictions generated from useful representations in ILB:

$$\mathcal{L}_{\text{sup}}\left(O, \hat{O}\right) = \frac{1}{N}\sum_{n=0}^{N}\left(O_n - \Phi_{spd}\left(\mathbf{h}_c\right)\right)^2 \tag{4}$$

Unlike the classification work [44] that uses uniform classification loss to eliminate the influence of irrelevant patterns, we design a useless loss $\mathcal{L}_{\text{usl}}$ for regression tasks. It pushes the useless representation to be unnecessary by minimizing its value to zero:

$$\mathcal{L}_{\text{usl}}\left(O, \hat{O}_{usl}\right) = \frac{1}{N}\sum_{n=0}^{N}\left(\Phi_{spd}\left(\mathbf{h}_u\right)\right)^2 \tag{5}$$

Then, the total loss for the causal learning strategy is:

$$\mathcal{L}_{\text{caus}} = \lambda_1\mathcal{L}_{\text{sup}} + \lambda_2\mathcal{L}_{\text{usl}} + \lambda_3\mathcal{L}_{int} \tag{6}$$

where λ_1, λ_2, λ_3 are hyperparameters. To this end, CityCAN leverages the intervention strategy in causal theory, guiding the causal framework to identify useful correlations among regions.

3.1.2 ***Global Local-Attention Encoder (GLAE).*** As mentioned in Section 3.1.1, we propose the GLAE, as shown in Fig. 3 (b), to capture the ST correlations in ILB and ULB. Since vehicles can travel at varying speeds, either quickly or slowly, in a city, it is essential to model both local and global ST correlations. Inspired by recent studies that employ temporal attentions [56, 71] to address short- and long-term temporal dependencies, GLAE extends the temporal attention [56, 71] to the spatio-temporal attention for citywide ST forecasting. A GLAE has M Calibration Attention (CA) Blocks, which include an auxiliary feature embedding, a local CA module (LCAM), a global CA module (GCAM), and a cropping layer. Since the architecture of GLAE is the same in ILB and ULB, we omit specific branch names in subsequent sections.

The auxiliary feature (Aux) embedding provides ST positional information and external factor information. It includes the ST positional (Pos) embedding and the external factor (Ext) embedding. To enhance the positional context, we introduce ST Pos embedding to encode the space-time location. It extends the canonical positional embedding [71] into ST format. For a superregion $B_l = B(n, t)$, the ST positional index of its location is $l = (n + tN_s)$, where $n \in N_s$ and $t \in T$ is the index of the space location and time location, respectively. We can obtain the ST Pos embedding $\boldsymbol{e}_{pos} \in \mathbb{R}^{L\times d_p}$ by applying the learnable positional embedding [46] to the ST positional index. To inject external factors into the latent features, we use learnable embedding layers to encode external factors and generate the Ext embedding $\boldsymbol{e}_{ext} \in \mathbb{R}^{L\times d_e}$. Then, we obtain the Aux embedding by concatenating the ST Pos embedding $\boldsymbol{e}_{pos}$ and Ext embedding $\boldsymbol{e}_{ext}$, i.e., $\mathbf{E} = \boldsymbol{e}_{pos}||\boldsymbol{e}_{ext} \in \mathbb{R}^{L\times d_h}$. After that, we obtain inputs for subsequent ST correlation modeling by adding the Aux embedding $\mathbf{E}$ to ST representations $\mathbf{h}$.

CA Module for Local & Global ST Learning The conventional attentions [22, 46, 71] cannot apply to ST data directly, as they primarily focus on temporal dimension, ignoring ST relationships. Citywide ST data has two crucial ST relationships: (1) future traffic conditions cannot affect past conditions; (2) spatially connected superregions have a higher influence on each other. To incorporate these relationships into attention operations, we proposed a Calibration Attention Module (CAM), whose core is a calibration attention (CA) operation. The CA operation works on calibrating the attention based on citywide ST relationships via two components:

- *ST influential mask* $\mathbf{M}_{st} \in \mathbb{R}^{L\times L}$ prevents future information leakage by setting the attention that represents influence from future time intervals with zero [46]. However, unlike the masks in prior works [25, 46], $\mathbf{M_{st}}$ is not a triangular matrix as the superregions are arranged by space-time location.
- *Spatial bias* $\mathbf{PE}_s \in \mathbb{R}^{L\times L}$ enhances spatial relationships by setting temporal influence to zero and repeating spatial influence with superregions' spatial relationships $\tilde{A}$.

Then, we revise the conventional attention operation to the calibration attention (CA) operation:

$$\text{Attention}\left(\mathbf{Q}, \mathbf{K}, \mathbf{V}\right) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} \cdot \mathbf{M}_{st} + \mathbf{PE}_s\right)\mathbf{V} \tag{7}$$

where $\mathbf{Q} \in \mathbb{R}^{L\times d_h}$, $\mathbf{K} \in \mathbb{R}^{L\times d_h}$, and $\mathbf{V} \in \mathbb{R}^{L\times d_h}$. Note that reshaping and broadcasting are needed to retain the ST positional index of superregions. In the CAM, H CA operations are performed to attend different ST patterns. Then, the output of the CAM is the aggregated ST representations $\vec{\mathbf{h}} \in \mathbb{R}^{L\times d_h}$, obtained by applying a layer normalization, a residual connection, and a fully connected feed-forward network on the concatenation of the H CA operations.

CAM can capture both local and global ST features owing to its attention-based design. To better learn local and global ST features, we use the CAM in two different ways: the local CAM (LCAM) and the global CAM (GCAM):

Local CAM (LCAM) captures local ST representations within a sliding window of size T_w. The total number of ST superregions in each window is $L_l = N_s T_w$. We apply CAM on these supperregions with their components $\mathbf{M}_{stl}, \mathbf{PE}_{sl} \in \mathbb{R}^{L_l\times L_l}$ (see calculations above). Given $T^{(m)}$ temporal intervals at m-th block, there are $Z^{(m)} = T - (m-1)(T_w - 1)$ sliding windows, resulting in $Z^{(m)} \cdot N_s \cdot T_w$ superregions. The index of m starts from 1. Since each superregion aggregates ST features from all other superregions, we let the LCAM only outputs the features of the last time interval for each sliding window, i.e., $\vec{\mathbf{h}}_l^{(m)} \in \mathbb{R}^{L^{(m)}\times d_h}$, where $L^{(m)} = N_s Z^{(m)} = N_s T^{(m)}$ is the total number of superregions in m-th block, and $T^{(m)}$ is:

$$T^{(m)} = \begin{cases} T - (m-1)(T_w - 1) & \text{if } (m-1)(T_w - 1) < T \\ T_w & \text{otherwise} \end{cases} \tag{8}$$

Global CAM (GCAM) learns the global ST features from all superregions across time and space. Similar to LCAM, we apply CAM on all ST superregions, i.e., $L^{(m)}$ superregions, with their corresponding calibration components $\mathbf{M}_{stg}, \mathbf{PE}_{sg} \in \mathbb{R}^{L^{(m)}\times L^{(m)}}$, to obtain the final output of GCAM $\vec{\mathbf{h}}_g^{(m)} \in \mathbb{R}^{L^{(m)}\times d_h}$.

Cropping Layer removes redundant features from the farthest superregions as traffic conditions are primarily influenced by the most adjacent time intervals. The redundant information resides in $\vec{\mathbf{h}}_g^{(m)}$ because each superregion has aggregated ST information from all other superregions in GCAM. Thus, at the last block M, we only use the superregions at the last temporal interval $T^{(M)}$,

i.e., $\mathbf{h} = \vec{\mathbf{h}}_g(:, -1) \in \mathbb{R}^{N_s \times d_h}$. Then, total number of superregions $L^{(m)}$ is updated to:

$$L^{(m)} = \begin{cases} N_s(T - m(T_w - 1)) & \text{if } m < T/(T_w - 1) \\ N_s & \text{if } m = M \\ N_s T_w & \text{otherwise} \end{cases} \tag{9}$$

3.2 Citywide Loss

Regions within a city, influenced by urban zoning and functionality, can be categorized into: (1) *significant regions*, characterized by frequent events and may require extra human interventions (e.g., traffic control in case of predicted accidents or pre-allocation of taxis for areas with high predicted demand). (2) *trivial regions*, which often have small or zero event numbers and do not require specific human interventions. Significant and trivial regions are non-evenly distributed in a city. To ensure effective interventions without wasted resources, the network should accurately predict targeted features for all regions in the city simultaneously. However, the causal loss (Eq. 6) emphasizes region-wise error, which can misalign predictions with the city's actual spatial distribution. To address this issue, we introduce an auxiliary loss, named *citywide loss*, to regularize the distribution between predictions and labels. Also, recognizing the heightened importance of significant regions, particularly in applications requiring costly human intervention, we first introduce a *calibration prior* to up-weight significant regions.

Calibration Prior leverages the citywide domain knowledge that a similar spatial distribution over time. This knowledge exists because traffic is influenced by the city's geography and semantics. Thus, we can identity the significant regions via a region prior P_r by summarizing the interested conditions features of each region over observed samples, i.e., training samples, and obtain the **calibration prior** P_c based on the region prior P_r:

$$\begin{aligned} P_r^{(n)} &= \sum_{i=1}^{\mathcal{I}} X_i^{(n)} \Big/ \max_{n \epsilon N}\Big(\sum_{i=1}^{\mathcal{I}} X_i^{(n)}\Big) \\ P_c^{(n)} &= \begin{cases} 1 + \exp(\, P_r^{(n)} - \tau) & \text{if } P_r^{(n)} > \tau \\ 1 & \text{otherwise} \end{cases} \end{aligned} \tag{10}$$

where $P_r \in \mathbb{R}^{N \times d_o}$, $\mathcal{I}$ is the total number of training samples, $n \in N$ is the index of spatial region, X refers to targeted traffic condition features, e.g., the features of traffic accident risk, taxi flow, crowd density, and τ is the *calibration parameter* that controls the selection of the most important significance regions.

Citywide Loss with Calibration Prior enables the network to generate the prediction distribution that can reflect the true citywide distribution, while penalizing the errors in significant regions more. We calculate the citywide loss based on the re-weighted cosine similarity:

$$\begin{aligned} O_c &= P_c \cdot O \qquad\qquad \hat{O}_c = P_c \cdot \hat{O} \\ \mathcal{L}_{\text{cwl}}\left(O_c, \hat{O}_c\right) &= 1 - \sum_{t=1}^{T_{pred}} \left(\frac{O_c^t \cdot \hat{O}_c^t}{\max\left(\left\|O_c^t\right\|_2 \cdot \left\|\hat{O}_c^t\right\|_2, \epsilon\right)} \right) \end{aligned} \tag{11}$$

where ϵ is a hyperparameter to avoid division by zero. The re-weighted cosine similarity is applied to all regions collectively for each time interval, thus can constrain the traffic condition spatial distribution. It also provides proper focus on each and every region, during training, as it re-weights the importance of the regions across the city. Note that the calibration prior is applied to both the predictions and labels, thus keeping the distribution.

3.3 Losses for CityCAN

The final loss for CityCAN contains two parts, i.e., *causal loss* in Eq. 6 and *citywide loss* (CWL) in Eq. 11:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{sup}} + \lambda_2 \mathcal{L}_{\text{usl}} + \lambda_3 \mathcal{L}_{int} + \lambda_4 \mathcal{L}_{cwl} \tag{12}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are hyperparameters. To this end, CityCAN can consider both the region-wise and citywide errors, and therefore ensures high predictive performance across all regions in the city.

4 EXPERIMENTS

4.1 Experimental Settings

4.1.1 ***Datasets.*** We evaluate CityCAN on four real-world datasets, i.e., NYC13 [47], BikeNYC [67], Chicago21 and Chicago22. NYC13 and BikeNYC are grid-based datasets with regular regions, while Chicago21 and Chicago22 dataset[1] contains irregular regions, better representing natural city divisions. Dataset details are in Table 1.

Table 1: The statistic of datasets.

Dataset	NYC13	BikeNYC	Chicago21	Chicago22
Time Span (mm/dd/yyyy)	01/01/2013 - 12/31/2013	04/01/2014 - 09/30/2014	01/01/2021- 12/31/2021	01/01/2022 - 12/31/2022
Time Interval	1 hour	1 hour	30 minutes	30 minutes
Region Size	(20, 20)	(16, 8)	77	77
Accident Severity	4	-	6	-
Weather Type	5	17	11	13

4.1.2 ***Tasks.*** We conduct experiments on three tasks, including the traffic accident risk forecasting, crowd flow forecasting, and crowd density forecasting:

Traffic accident risk forecasting task: we follow the existing works [6, 47], not only to predict the occurrence of traffic accidents, but also to estimate the risk value. The risk value should reflect both the frequency and severity of traffic accidents in the region, and thus it is defined as the sum of each traffic accident's severity within a region. In the experiments, we forecast traffic accident risk conditions for the next time interval ($T_{pred} = 1$) given historical observations of 6-time intervals ($T = 6$).

Crowd flow forecasting task and crowd density task: we follow existing works [27, 63] to predict the crowd inflow/outflow and crowd density value for all regions in the city, respectively. In the experiments, we predict crowd flow conditions and crowd density conditions for next 6 time intervals ($T_{pred} = 6$) given historical observations of 6-time intervals ($T = 6$).

4.1.3 ***Evaluation Metrics.*** We follow the previous studies [47, 50] to evaluate our model with two metrics: **Mean Absolute Error** (MAE) and **Root Mean Squared Errors** (RMSE). Additionally, towards more comprehensive evaluations of traffic accident risk forecasting, we use **F1 score**, **F1@20**, **F1@30** to present the ability

[1] https://data.cityofchicago.org/

of the model to indicate regions with risk, where F1@K denotes the F1 score for top K regions with high accident values.

4.1.4 ***Implementation Details.*** Our model is trained on a single GTX 2080 Ti using Adam optimizer with a learning rate of 0.001. We set region reduction parameter r to 4, number of blocks M to 4, feature dimension d_h to 128, number of multi-heads H to 4. We balance the region-wise and city-wise losses and set $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ to 0.4, 0.05, 0.05, and 0.5, respectively. The calibration parameter τ is dataset-specific. We adopt an early-stop strategy with a maximum of 150 epochs for all experiments. For the data partitioning, we used the last 8 weeks as the test set, the preceding 4 weeks as the validation set, and the remaining data as the training set.

4.1.5 ***Baselines.*** To fully demonstrate the effectiveness of CityCAN across different tasks, we adopt the following baselines that are specifically designed for each task:

- **Classical methods**: HA [4] averages the historical traffic conditions of the same time slot given the past observed segments.
- **Traffic accident risk prediction**: we select three popular methods, i.e., SDAE [6], SDCAE [5], and GSNet [47], and two general ST models, i.e., STGCN [63] and GWNET [58], as baselines.
- **Crowd flow & crowd density prediction**: we select six strong models for comparisons, including STGCN [63], DCRNN [27], GWNET [58], AGCRN [2], MTGNN [57], and GMSDR [31].

4.2 Experimental Results & Analysis

4.2.1 ***Traffic Accident Risk Forecasting.*** Table 2 shows the prediction results of baselines and our model for traffic accident risk on two datasets. Our model consistently surpasses the baselines on all datasets in terms of accuracy for risk value and F1 for accident indication. Note that most recent works, i.e., SDCAE and GSNet, cannot adapt to the Chicago21 dataset. This is because these models, designed for regular regions, employ CNNs for spatial capturing. However, the Chicago21 dataset reflects the natural community divisions of a city, containing irregular regions. These regions have a non-Euclidean structure, which cannot be modeled using CNNs.

Table 2: Model comparisons on the NYC13 and Chicago21 datasets for traffic accident forecasting, where - denotes the model cannot be applied on the dataset.

Dataset	Method	MAE↓	RMSE↓	F1↑	F1@30↑	F1@20↑
NYC13	HA	0.05	0.28	11.10%	10.98%	11.02%
	SDAE	0.07	0.48	5.44%	32.38%	37.45%
	SDCAE	0.10	0.70	5.89%	52.88%	60.17%
	GSNet	0.04	0.25	7.19%	24.45%	26.09%
	STGCN	0.05	0.28	3.97%	4.04%	4.36%
	GWNET	0.04	0.27	3.01%	3.11%	3.46%
	Ours	**0.03**	**0.23**	**21.44%**	**62.07%**	**68.95%**
Chicago21	HA	0.17	0.84	12.36%	16.68%	18.51%
	SDAE	0.17	0.82	11.40%	19.89%	23.28%
	SDCAE	-	-	-	-	-
	GSNet	-	-	-	-	-
	STGCN	0.16	0.83	9.44 %	9.53%	9.67%
	GWNET	0.18	0.86	3.71%	3.73%	3.79%
	Ours	**0.13**	**0.72**	**18.64%**	**24.45%**	**27.49%**

From the results, we can observe that: (1) Our model outperforms baselines by a large margin on accident indication. Specifically, it brings 2.07 times higher citywide F1 on average, demonstrating its superior ability to identify regions with/without accident incidents. (2) Baselines designed for traffic accident risk forecasting (i.e., SDAE, SDCAE, and GSNet) perform better than general ST forecasting models (i.e., STGCN and GWNET). This is because that accidents are rare events, and general ST models, focusing on ST modeling, fail to consider the sparse data inherent in this task. (3) Surprisingly, HA outperforms all baselines on city-wide F1 scores. We conjecture that deep models focus on significant regions and neglecting trivial ones, thereby failing to identify trivial regions and leading to lower city-wide F1 scores. On the other hand, HA only considers historical observations for each region, avoiding this issue. Our model, adopting the citywide loss, considers the prediction distribution and places proper focus on each region, resulting in the best performance. (3) Our model achieves the lowest MAE and RMSE error, suggesting that it can generate more accurate risk values, since it generates predictions based on the useful correlations that truly impact the future condition.

Table 3: Model comparisons for crowd flow prediction on BikeNYC dataset and Chicago21 dataset, and crowd density prediction on Chicago22 dataset.

Method	BikeNYC		Chicago21		Chicago22	
	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓
HA	5.18	9.19	1.26	3.93	3.17	11.82
STGCN	3.69	7.69	1.10	3.08	2.48	7.71
DCRNN	4.49	7.66	1.40	2.87	2.57	7.26
GWNET	4.49	8.08	1.30	2.66	2.63	6.41
AGCRN	5.04	8.74	1.36	2.84	2.84	6.51
MTGNN	4.01	8.02	1.35	2.77	2.60	7.03
GMSDR	5.19	8.58	1.31	2.67	2.37	6.19
Ours	**3.57**	**7.31**	**1.08**	**2.61**	**2.28**	**6.09**

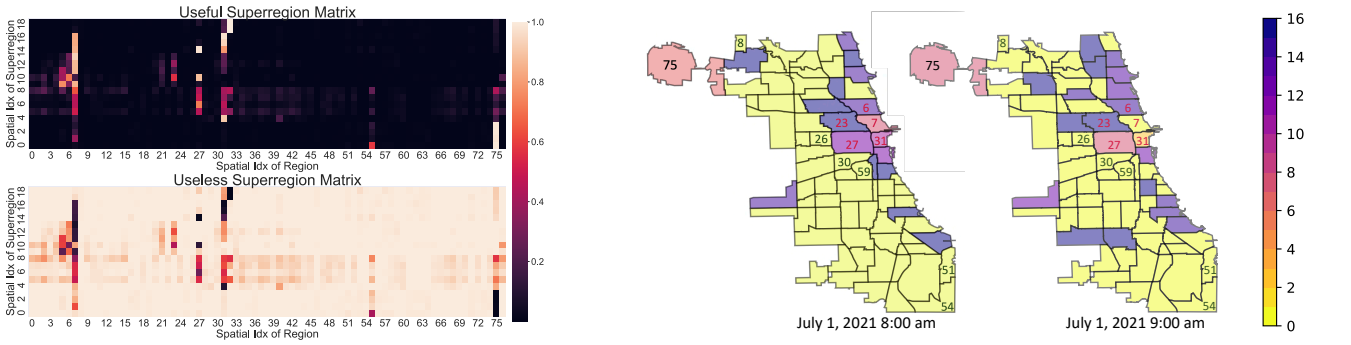
4.2.2 ***Crowd Flow Forecasting & Crowd Density Forecasting.*** Table 3 shows the results of baselines and our model for crowd flow forecasting on BikeNYC and Chicago21 dataset, and crowd density forecasting on Chicago22 dataset. Compared to the accident risk forecasting task, the data in these two tasks are not sparse. The results show that our model consistently outperforms existing methods on all metrics. Specifically, it reduces MAE error by 17.78% and RMSE error by 14.47% on average over three datasets. It demonstrates that our proposed model is a general model which can achieve better performance on various citywide tasks.

Table 4: Ablation studies of CityCAN for crowd density prediction on Chicago22 dataset.

Method	w/o C	w/o UIL	w/o UL	w/o IL	w/o CWL
MAE	2.31	2.41	2.34	2.45	2.27
RMSE	8.71	6.25	6.14	6.35	6.25
Method	**w/o CP**	**w/o CL**	**w/o LCAM**	**w/o GCAM**	**CityCAN**
MAE	**2.26**	2.58	2.30	2.88	2.28
RMSE	6.16	6.59	6.58	7.33	**6.09**

4.3 Ablation Study

Table 4 details the effectiveness of each component in CityCAN. **w/o C** lacks the causal framework, which adopts a single invariant learning module. **w/o UIL** it the model without both useless loss



(a) Superregion metrices (b) Crowd inflow

Figure 4: (a) A visualization of superregion metrices for crowd flow forecasting on Chicago21 dataset. (b) A visualization of crowd inflow on Chicago21 dataset, where the idx (index) in each region refers to its spatial index.

(Eq. 5) and intervention loss (Eq. 3). **w/o UL** excludes the useless loss. **w/o IL** omits the intervention loss. **w/o CWL** is the model without the citywide loss (Eq. 11). **w/o CP** does not have the calibration prior within the citywide loss. **w/o CL** removes the cropping layer. **w/o LCAM** and **w/o GCAM** are models without the LCAM and GCAM modules, respectively.

Table 4 reveals: (1) Causal learning strategy improves performance, validating the effectiveness of identifying useful correlations. (2) Excluding useless loss or intervention loss hurts the model performance, indicating useless correlations do exist and misleads the network. These two losses must work together to achieve causal learning as useless loss ensures zero influence of useless features, while intervention loss guarantees invariant results after adding useless features. (3) Omitting citywide loss degrades performance, demonstrating the importance of considering all city regions. (4) Higher RMSE in **w/o CP** indicates that the calibration prior can provide useful domain knowledge to enhance performance in regions with high condition values. Although applying the calibration prior tends to focus more on RMSE, resulting in a slight increase in MAE, it is particularly useful in scenarios where high condition values are of high interest, such as traffic accident risk. Its influence can easily be removed by setting the calibration parameter to 1. (5) The inferior performance of **w/o CL** highlights that removing redundant information can make the model focus on the most important features. (6) **w/o LCAM** and **w/o GCAM** show inferior performance, demonstrating that capturing the local and global ST correlations is necessary for citywide ST forecasting.

4.4 Visualization

4.4.1 ***Superregion Matrices.*** Fig. 4 displays the two superregion matrices, i.e., useful supperregion matrix and useless superregion matrix, for crowd flow forecasting on Chicago21 dataset. From Fig. 4, we can observe that: (1) The two metrices are complementary to one another, which disentangle useful correlations from useless ones successfully. (2) Useful correlations are discovered because regions with useful correlations are grouped into one superregion. For example, central business district (CBD) regions like Region 7 and Region 31, along with residential regions such as Region 27, are assigned to the same superregion (e.g., Superregion 14). This grouping indicates their correlation and aligns with known city studies insights. (3) Certain rural regions, e.g., Region 8 and Region

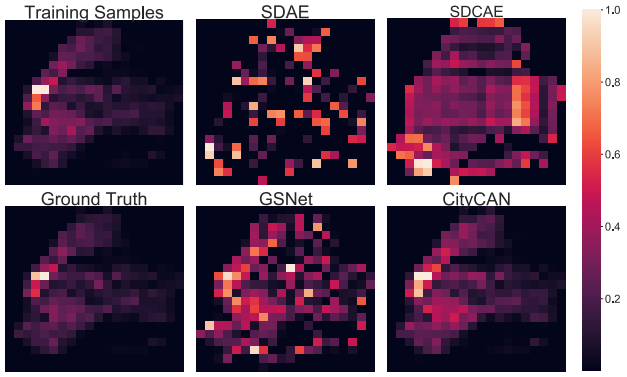
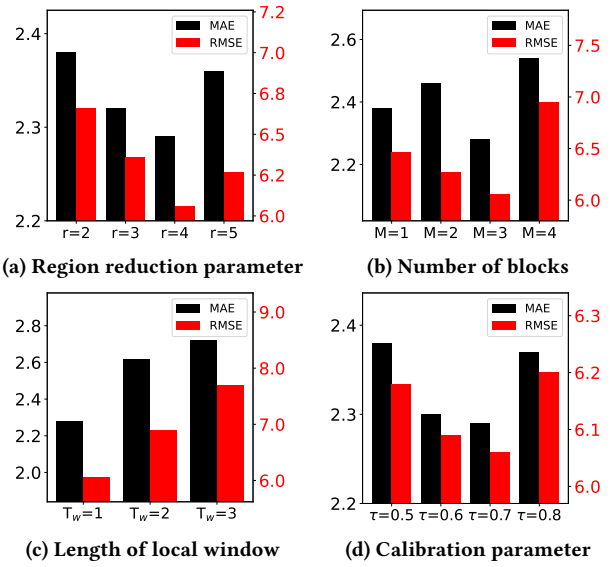


Figure 5: A visualization of the citywide spatial distribution of traffic accidents in the NYC13 dataset, along with the forecasting results of various methods.

51, have fewer correlations with other regions and thus have higher weights in the useless superregion matrix. (4) Our model allows each region to be assigned to multiple superregions based on its correlations with different regions. For instance, Region 7 is also assigned to Superregion 9, as it also has correlations with Region 5 and Region 6. One region's weight to different superregions varies according to its importance within different correlations.

4.4.2 ***Citywide Distribution.*** Fig. 5 shows the distribution of the citywide data on the NYC13 dataset, which is obtained by averaging the traffic accident risk values over the temporal dimension and normalizing the traffic accident risk values over the spatial dimension. The visualization of training samples is derived from the training set, while the ground truth represents the visualization of the ground truth of forecasting in the test set. These two visualizations share a similar distribution, validating our assumption that the citywide distribution does not vary dramatically between the training and test set. The forecast visualizations produced by SDAE, SDCAE, and GSNet illustrate these models' limitations in generating good predictions that align with the actual citywide distribution. It is because they adopt region-wise losses, which neglect the citywide distribution. Especially, they fail to identify trivial regions, which consistently present zero values over time, and may lead to inefficient resource allocation. CityCAN considers both region-wise and citywide errors in forecasting, and thus can



(a) Region reduction parameter (b) Number of blocks

(c) Length of local window (d) Calibration parameter

Figure 6: Effects of hyperparameters on Chicago22 dataset in terms of MAE and RMSE.

successfully recognize significant and trivial regions and achieve good forecasting results for all regions in the city.

4.5 Effects of Hyperparameters

In Fig. 6, we study the effects of hyperparameters in CityCAN over Chicago22 for crowd density forecasting. From the results, we observe: (1) CityCAN achieves the lowest RMSE error when the region reduction parameter $r = 4$. This is because a higher reduction rate results in fewer superregions, which allows the network to obtain summarized features from original correlated regions and eliminates some redundant features. However, if the reduction rate is too high, it may contain many useless correlations, which negatively impact performance. (2) CityCAN achieves the highest performance when it contains $M = 3$ CA blocks as reducing/increasing the number of blocks may lead to underfitting/overfitting issues. (3) Increasing the local window length negatively impacts performance, as it is important to consider temporal information from each time interval for short-term forecasting. (4) CityCAN performs best when $\tau = 0.7$, particularly in terms of RMSE, because it gives suitable weights to the significant regions that have higher values.

5 RELATED WORK

Citywide Spatio-Temporal Forecasting is a crucial task for ITS and has attracted much attention over the years. Recent works have explored spatio-temporal (ST) networks for various citywide tasks, such as traffic accident prediction [37, 49, 65, 76], traffic flow prediction [13, 20, 26, 33, 51], traffic speed prediction [9, 70], taxi demand prediction [1, 62, 78], etc. They have achieved superior performance over traditional statistical models, like k-nearest neighbor [35] and ARIMA [45, 54] thanks to their ability to model complex nonlinear ST correlations. More recent works [29, 43, 48] suggest that jointly learning the spatial and temporal dependencies enhances prediction performance. However, they still face challenges in considering the global ST correlations between the irregular regions simultaneously. Meanwhile, attention-based models [56, 71] have shown success in learning global dynamic dependencies on temporal forecasting. However, they focus on long-term multivariate time-series and efficient attention mechanisms [23, 25, 32, 36, 52, 72], ignoring spatial correlations and domain knowledge, and therefore cannot be applied to citywide ST forecasting directly. Also, in citywide forecasting, citywide distribution is relatively under-explored. Although some works [15, 24, 47, 55, 75] have studied zero-inflated data that are distributed sparsely in the city. They focus on region-wise optimization, which results in producing skewed predictions that cannot align with the citywide distribution. In this work, we propose a novel attention-based ST encoder that incorporates citywide domain knowledge in a casual framework and a citywide loss to constrain the prediction distribution for better ST modeling.

Causal Learning [38, 39] enables the deep learning models with the ability to eliminate spurious correlations, leading to improved performance in various tasks. For example, CONTA [66] removes non-causal associations between image pixels and labels via the backdoor adjustment in image segmentation tasks. Liu et al. [34] learns the causal invariance of the motion representations by disentangling the physical laws, style confounders, and non-casual features for better motion prediction. CAL [44] boosts graph classification performance by applying causal interventions on representation level. STNSCM [10] analyze the causal relationship between input data and contextual conditions. Different from them, we propose a causal attention network that removes the useless correlations that exist in ST data for citywide regression. There are some concurrent studies on causal learning for ST data [59, 74].

6 CONCLUSION AND FUTURE WORK

In this paper, we proposed CityCAN, novel network for citywide ST forecasting. Leveraging the causal theory, we design a causal framework for citywide ST forecasting that applies implicit interventions at the latent level, enabling CityCAN to learn useful regional correlations. To jointly capture the ST correlations for both regular and irregular regions, we also introduce a Global-Local Attention Encoder in CityCAN. It captures both the local and global ST correlations with a calibrated attention mechanism for better ST modeling. We then proposed a citywide loss, which considers the citywide distribution between the predicted and real conditions, to enable CityCAN to accurately predict the targeted features for all regions in the city at once. Extensive experimental results and analyses verified the effectiveness of CityCAN. CityCAN is not limited to citywide ST forecasting. In the future, we will evaluate it on other ST tasks, such as crime prediction. We will also exploit the different architectures for invariant learning and useless learning to reduce computational costs.

7 ACKNOWLEDGEMENT

This research is supported by the National Research Foundation, Singapore under its Industry Alignment Fund – Pre-positioning (IAF-PP) Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

REFERENCES

[1] Lei Bai, Lina Yao, Salil S. Kanhere, Xianzhi Wang, and Quan Z. Sheng. 2019. STG2Seq: Spatial-Temporal Graph to Sequence Model for Multi-step Passenger Demand Forecasting. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, Sarit Kraus (Ed.). ijcai.org, 1981–1987. https://doi.org/10.24963/ijcai.2019/274

[2] Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. 2020. Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in neural information processing systems* 33 (2020), 17804–17815.

[3] Jie Bao, Pan Liu, and Satish V Ukkusuri. 2019. A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data. *Accident Analysis & Prevention* 122 (2019), 239–254.

[4] Peter J Brockwell, Peter J Brockwell, Richard A Davis, and Richard A Davis. 2016. *Introduction to time series and forecasting*. Springer.

[5] Chao Chen, Xiaoliang Fan, Chuanpan Zheng, Lujing Xiao, Ming Cheng, and Cheng Wang. 2018. Sdcae: Stack denoising convolutional autoencoder model for accident risk prediction via traffic big data. In *2018 Sixth International Conference on Advanced Cloud and Big Data (CBD)*. IEEE, 328–333.

[6] Quanjun Chen, Xuan Song, Harutoshi Yamada, and Ryosuke Shibasaki. 2016. Learning deep representation from big and heterogeneous data for traffic accident inference. In *Thirtieth AAAI conference on artificial intelligence*.

[7] Yile Chen, Xiucheng Li, Gao Cong, Cheng Long, Zhifeng Bao, Shang Liu, Wanli Gu, and Fuzheng Zhang. 2021. Points-of-Interest Relationship Inference with Spatial-enriched Graph Neural Networks. *Proc. VLDB Endow.* (2021), 504–512.

[8] Yile Chen, Cheng Long, Gao Cong, and Chenliang Li. 2020. Context-aware deep model for joint mobility and time prediction. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 106–114.

[9] Jeongwhan Choi, Hwangyong Choi, Jeehyun Hwang, and Noseong Park. 2022. Graph neural controlled differential equations for traffic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 6367–6374.

[10] Pan Deng, Yu Zhao, Junting Liu, Xiaofeng Jia, and Mulan Wang. 2023. Spatio-Temporal Neural Structural Causal Models for Bike Flow Prediction. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, Brian Williams, Yiling Chen, and Jennifer Neville (Eds.). AAAI Press, 4242–4249.

[11] Kaiqun Fu, Taoran Ji, Liang Zhao, and Chang-Tien Lu. 2019. Titan: A spatiotemporal feature learning framework for traffic incident duration prediction. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 329–338.

[12] Xu Geng, Yaguang Li, Leye Wang, Lingyu Zhang, Qiang Yang, Jieping Ye, and Yan Liu. 2019. Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting. In *Proceedings of the AAAI conference on artificial intelligence*. 3656–3663.

[13] Shengnan Guo, Youfang Lin, Huaiyu Wan, Xiucheng Li, and Gao Cong. 2021. Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting. *IEEE Transactions on Knowledge and Data Engineering* 34, 11 (2021), 5415–5428.

[14] Liangzhe Han, Bowen Du, Leilei Sun, Yanjie Fu, Yisheng Lv, and Hui Xiong. 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.

[15] Chao Huang, Junbo Zhang, Yu Zheng, and Nitesh V Chawla. 2018. DeepCrime: Attentive hierarchical recurrent networks for crime prediction. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 1423–1432.

[16] Benedikt Jäger, Michael Wittmann, and Markus Lienkamp. 2016. Analyzing and modeling a City's spatiotemporal taxi supply and demand: A case study for Munich. *Journal of Traffic and Logistics Engineering* 4, 2 (2016).

[17] Jiahao Ji, Jingyuan Wang, Zhe Jiang, Jiawei Jiang, and Hu Zhang. 2022. STDEN: Towards Physics-Guided Neural Networks for Traffic Flow Prediction. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 4048–4056.

[18] Jiahao Ji, Jingyuan Wang, Zhe Jiang, Jingtian Ma, and Hu Zhang. 2020. Interpretable spatiotemporal deep learning model for traffic flow prediction based on potential energy fields. In *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1076–1081.

[19] Renhe Jiang, Zekun Cai, Zhaonan Wang, Chuang Yang, Zipei Fan, Quanjun Chen, Kota Tsubouchi, Xuan Song, and Ryosuke Shibasaki. 2021. Deepcrowd: A deep model for large-scale citywide crowd density and flow prediction. *IEEE Transactions on Knowledge and Data Engineering* 35, 1 (2021), 276–290.

[20] Guangyin Jin, Fuxian Li, Jinlei Zhang, Mudan Wang, and Jincai Huang. 2022. Automated Dilated Spatio-Temporal Synchronous Graph Modeling for Traffic Prediction. *IEEE Transactions on Intelligent Transportation Systems* (2022).

[21] Guangyin Jin, Yuxuan Liang, Yuchen Fang, Jincai Huang, Junbo Zhang, and Yu Zheng. 2023. Spatio-temporal graph neural networks for predictive learning in urban computing: A survey. *arXiv preprint arXiv:2303.14483* (2023).

[22] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*. 4171–4186.

[23] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The Efficient Transformer. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

[24] Xiaoliang Lei, Hao Mei, Bin Shi, and Hua Wei. 2022. Modeling Network-level Traffic Flow Transitions on Sparse Data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 835–845.

[25] Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. 2019. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Advances in neural information processing systems* 32 (2019).

[26] Ting Li, Junbo Zhang, Kainan Bao, Yuxuan Liang, Yexin Li, and Yu Zheng. 2020. Autost: Efficient Neural Architecture Search for Spatio-Temporal Prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 794–802.

[27] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2018. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

[28] Yuxuan Liang, Kun Ouyang, Junkai Sun, Yiwei Wang, Junbo Zhang, Yu Zheng, David Rosenblum, and Roger Zimmermann. 2021. Fine-grained urban flow prediction. In *Proceedings of the Web Conference 2021*. 1833–1845.

[29] Yuxuan Liang, Kun Ouyang, Yiwei Wang, Ye Liu, Junbo Zhang, Yu Zheng, and David S. Rosenblum. 2020. Revisiting Convolutional Neural Networks for Citywide Crowd Flow Analytics. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2020, Ghent, Belgium, September 14-18, 2020, Proceedings, Part I*, Vol. 12457. Springer, 578–594.

[30] Ziqian Lin, Jie Feng, Ziyang Lu, Yong Li, and Depeng Jin. 2019. Deepstn+: Context-aware spatial-temporal neural network for crowd flow prediction in metropolis. In *Proceedings of the AAAI conference on artificial intelligence*. 1020–1027.

[31] Dachuan Liu, Jin Wang, Shuo Shang, and Peng Han. 2022. Msdr: Multi-step dependency relation networks for spatial temporal forecasting. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1042–1050.

[32] Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. 2021. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International Conference on Learning Representations*.

[33] Xu Liu, Yutong Xia, Yuxuan Liang, Junfeng Hu, Yiwei Wang, Lei Bai, Chao Huang, Zhenguang Liu, Bryan Hooi, and Roger Zimmermann. 2023. LargeST: A Benchmark Dataset for Large-Scale Traffic Forecasting. *arXiv preprint arXiv:2306.08259* (2023).

[34] Yuejiang Liu, Riccardo Cadei, Jonas Schweizer, Sherwin Bahmani, and Alexandre Alahi. 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.

[35] Yisheng Lv, Shuming Tang, and Hongxia Zhao. 2009. Real-time highway traffic accident prediction based on the k-nearest neighbor method. In *2009 international conference on measuring technology and mechatronics automation*, Vol. 3. IEEE, 547–550.

[36] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *International Conference on Learning Representations*.

[37] Muhammad Shalihin Othman, Remya K Padinjarapat, Chengxin Wang, Nimal R Arunachalam, and Gary Tan. 2023. Real-Time Simulation Framework with Traffic Incident Prediction: A Singapore Case Study. In *2023 IEEE/ACM 27th International Symposium on Distributed Simulation and Real Time Applications (DS-RT)*. IEEE Computer Society, 84–90.

[38] Judea Pearl. 2014. Interpretation and identification of causal mediation. *Psychological methods* 19, 4 (2014), 459.

[39] Judea Pearl et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress* 19, 2 (2000).

[40] Xinwu Qian and Satish V Ukkusuri. 2015. Spatial variation of the urban taxi ridership using GPS data. *Applied geography* 59 (2015), 31–42.

[41] Khaled Saleh, Artur Grigorev, and Adriana-Simona Mihaita. 2022. Traffic Accident Risk Forecasting using Contextual Vision Transformers. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2086–2092.

[42] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems* 28 (2015).

[43] Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. 2020. Spatial-temporal synchronous graph convolutional networks: A new framework for

spatial-temporal network data forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 914–921.

[44] Yongduo Sui, Xiang Wang, Jiancan Wu, Min Lin, Xiangnan He, and Tat-Seng Chua. 2022. Causal attention for interpretable and generalizable graph classification. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1696–1705.

[45] Quang Thanh Tran, Zhihua Ma, Hengchao Li, Li Hao, and Quang Khai Trinh. 2015. A multiplicative seasonal ARIMA/GARCH model in EVN traffic prediction. *International Journal of Communications, Network and System Sciences* 8, 4 (2015), 43.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[47] Beibei Wang, Youfang Lin, Shengnan Guo, and Huaiyu Wan. 2021. GSNet: Learning spatial-temporal correlations from geographical and semantic aspects for traffic accident risk forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 4402–4409.

[48] Chengxin Wang, Yuxuan Liang, and Gary Tan. 2022. Periodic residual learning for crowd flow forecasting. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*. 1–10.

[49] Chengxin Wang and Gary Tan. 2023. Spatio-Temporal Forecasting for Traffic Simulation Framework. In *2023 IEEE/ACM 27th International Symposium on Distributed Simulation and Real Time Applications (DS-RT)*. IEEE Computer Society, 109–110.

[50] Jingyuan Wang, Jiawei Jiang, Wenjun Jiang, Chao Li, and Wayne Xin Zhao. 2021. Libcity: An open library for traffic prediction. In *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*. 145–148.

[51] Leye Wang, Xu Geng, Xiaojuan Ma, Feng Liu, and Qiang Yang. 2019. Cross-City Transfer Learning for Deep Spatio-Temporal Prediction. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*. 1893–1899.

[52] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768* (2020).

[53] Zhaonan Wang, Renhe Jiang, Zekun Cai, Zipei Fan, Xin Liu, Kyoung-Sook Kim, Xuan Song, and Ryosuke Shibasaki. 2021. Spatio-temporal-categorical graph neural networks for fine-grained multi-incident co-prediction. In *Proceedings of the 30th ACM international conference on information & knowledge management*. 2060–2069.

[54] Billy M Williams. 1999. *Modeling and forecasting vehicular traffic flow as a seasonal stochastic time series process*. University of Virginia.

[55] Tyler Wilson, Andrew McDonald, Asadullah Hill Galib, Pang-Ning Tan, and Lifeng Luo. 2022. Beyond Point Prediction: Capturing Zero-Inflated & Heavy-Tailed Spatiotemporal Data with Deep Extreme Mixture Models. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2020–2028.

[56] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems* 34 (2021), 22419–22430.

[57] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. 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*. 753–763.

[58] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. 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*. ijcai.org, 1907–1913.

[59] Yutong Xia, Yuxuan Liang, Haomin Wen, Xu Liu, Kun Wang, Zhengyang Zhou, and Roger Zimmermann. 2023. Deciphering Spatio-Temporal Graph Forecasting: A Causal Lens and Treatment. *arXiv preprint arXiv:2309.13378* (2023).

[60] Huaxiu Yao, Xianfeng Tang, Hua Wei, Guanjie Zheng, and Zhenhui Li. 2019. Revisiting Spatial-Temporal Similarity: A Deep Learning Framework for Traffic Prediction. In *2019 AAAI Conference on Artificial Intelligence (AAAI'19)*.

[61] Huaxiu Yao, Fei Wu, Jintao Ke, Xianfeng Tang, Yitian Jia, Siyu Lu, Pinghua Gong, Jieping Ye, and Zhenhui Li. 2018. Deep multi-view spatial-temporal network for taxi demand prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[62] Junchen Ye, Leilei Sun, Bowen Du, Yanjie Fu, and Hui Xiong. 2021. Coupled layer-wise graph convolution for transportation demand prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 4617–4625.

[63] Bing Yu, Haoteng Yin, and Zhanxing Zhu. 2018. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, Jérôme Lang (Ed.). ijcai.org, 3634–3640.

[64] Haitao Yu and Zhong-Ren Peng. 2019. Exploring the spatial variation of ridesourcing demand and its relationship to built environment and socioeconomic factors with the geographically weighted Poisson regression. *Journal of Transport Geography* 75 (2019), 147–163.

[65] Zhuoning Yuan, Xun Zhou, and Tianbao Yang. 2018. Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 984–992.

[66] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. 2020. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems* 33 (2020), 655–666.

[67] Junbo Zhang, Yu Zheng, and Dekang Qi. 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Thirty-first AAAI conference on artificial intelligence*.

[68] Xiyue Zhang, Chao Huang, Yong Xu, Lianghao Xia, Peng Dai, Liefeng Bo, Junbo Zhang, and Yu Zheng. 2021. Traffic Flow Forecasting with Spatial-Temporal Graph Diffusion Network. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*. AAAI Press, 15008–15015.

[69] Wei Zhao, Shiqi Zhang, Bei Wang, and Bing Zhou. 2023. Spatio-temporal causal graph attention network for traffic flow prediction in intelligent transportation systems. *PeerJ Computer Science* 9 (2023), e1484.

[70] Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. 2020. Gman: A graph multi-attention network for traffic prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 1234–1241.

[71] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11106–11115.

[72] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. 2022. FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.). PMLR, 27268–27286.

[73] Yirong Zhou, Jun Li, Hao Chen, Ye Wu, Jiangjiang Wu, and Luo Chen. 2020. A spatiotemporal attention mechanism-based model for multi-step citywide passenger demand prediction. *Information Sciences* 513 (2020), 372–385.

[74] Zhengyang Zhou, Qihe Huang, Kuo Yang, Kun Wang, Xu Wang, Yudong Zhang, Yuxuan Liang, and Yang Wang. 2023. Maintaining the Status Quo: Capturing Invariant Relations for OOD Spatiotemporal Learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[75] Zhengyang Zhou, Yang Wang, Xike Xie, Lianliang Chen, and Hengchang Liu. 2020. RiskOracle: a minute-level citywide traffic accident forecasting framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 1258–1265.

[76] Zhengyang Zhou, Yang Wang, Xike Xie, Lianliang Chen, and Chaochao Zhu. 2020. Foresee urban sparse traffic accidents: A spatiotemporal multi-granularity perspective. *IEEE Transactions on Knowledge and Data Engineering* (2020).

[77] Pengyu Zhu, Jie Huang, Jiaoe Wang, Yu Liu, Jiarong Li, Mingshu Wang, and Wei Qiang. 2022. Understanding taxi ridership with spatial spillover effects and temporal dynamics. *Cities* 125 (2022), 103637.

[78] Ali Zonoozi, Jung-jae Kim, Xiao-Li Li, and Gao Cong. 2018. Periodic-CRN: A Convolutional Recurrent Model for Crowd Density Prediction with Recurring Periodic Patterns.. In *IJCAI*. 3732–3738.