# CityCAN: Causal Attention Network for Citywide Spatio-Temporal Forecasting

# CityCAN: Causal Attention Network for Citywide Spatio-Temporal Forecasting

Chengxin Wang

Chengxin Wang

cwang@comp.nus.edu.sg

National University of Singapore

National University of Singapore

Singapore

Singapore

Yuxuan Liang

Yuxuan Liang

yuxliang@outlook.com

The Hong Kong University of Science

The Hong Kong University of Science

and Technology (Guangzhou)

and Technology (Guangzhou)

China

China

Gary Tan

Gary Tan

gtan@comp.nus.edu.sg

National University of Singapore

National University of Singapore

# ABSTRACT

# 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.

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

# CCS CONCEPTS

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

# KEYWORDS

# KEYWORDS

Spatio-temporal Data Mining, Causal Intervention, Attention

Spatio-temporal Data Mining, Causal Intervention, Attention

# ACM Reference Format:

# 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

王成新，梁宇轩，陈国添。2024。CityCAN：面向城市范围时空预测的因果注意力网络。在2024年第17届ACM国际网络搜索与数据挖掘会议（WSDM ’24）论文集，2024年3月4-8日，墨西哥梅里达。ACM，纽约州纽约市，美国，10页。https://doi.org/10.1145/3616855.3635764

# 1 INTRODUCTION

# 1 引言

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

为了构建智能交通系统（ITS），城市中广泛部署了众多传感器来捕捉交通状况 [21]，产生了大量的时空（ST）数据，如图1（a）所示。预测城市范围的ST数据，如交通事故、人流量和人群密度，对ITS的发展至关重要。

Singapore

新加坡

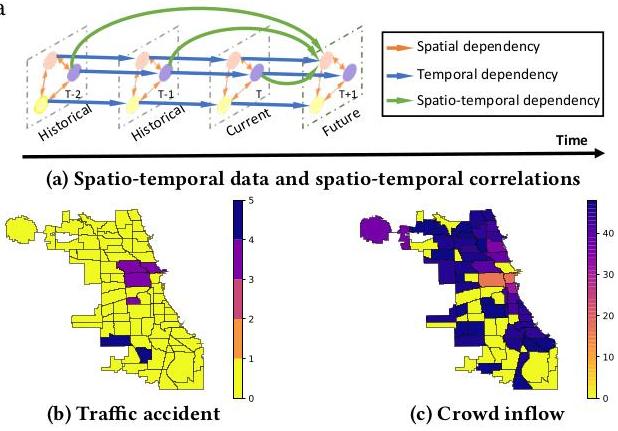


Figure 1: (a) A illustration of the spatio-temporal data and three types of dependency. (b-c) A visualization of citywide traffic accidents and crowd inflow in Chicago on July 1, 2021 at 14:00. Traffic conditions are influenced by urban zoning.

图1：（a）时空数据及三种依赖关系的示意图。（b-c）2021年7月1日14:00在芝加哥的城市范围交通事故和人群流入的可视化。交通状况受到城市规划的影响。

flows, and crowd densities, is crucial for ITS development. It can facilitate various urban applications, such as assisting transportation managers in mitigating accidents , guiding car-sharing companies in vehicle allocation , and aiding drivers in selecting optimal routes .

这些预测对于ITS的发展至关重要。它们可以促进各种城市应用，例如协助交通管理者减轻交通事故 ，指导共享汽车公司在车辆分配上 ，以及帮助驾驶者选择最佳路线 。

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 spatiotemporal (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) or graph neural networks (GNNs) [14, 58], and exploit temporal dependencies with recurrent mechanisms (RNNs) or attention mechanisms . 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:

城市范围共享单车（ST）数据的一个关键特征是时空（ST）相关性，这在图1（a）中有所说明。具体来说，目标区域的情况受到三种依赖关系的影响：空间依赖（由橙色线表示）、时间依赖（蓝色线）以及时空（ST）依赖（绿色线）。在大数据时代，研究人员提出了许多数据驱动方法，特别是深度学习方法，来捕捉这些ST相关性。大多数工作[47, 60]分别处理空间依赖和时间依赖，忽略了直接的ST依赖。他们通常通过卷积神经网络（CNNs） 或图神经网络（GNNs）[14, 58]来捕捉空间依赖，利用循环机制（RNNs） 或注意力机制 来开发时间依赖。为了充分利用ST相关性，更近的方法通过局部ST图[43]、金字塔CNNs[29]或ST增强机制[48]同时建模空间依赖和时间依赖。尽管在ST建模方面取得了最近的进展，但在预测城市范围ST数据时仍存在两个主要挑战：

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 . 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 networks to learn spatial correlations, often leading to an overgeneralized 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.

挑战一：如何识别跨时间区域之间的有用相关性？研究表明，城市规划分区和功能会影响全城的时空相关性 。为了将城市规划功能融入全城预测，先前研究 [28, 30] 将地理特征（例如，兴趣点）作为时空网络的辅助输入。然而，这些方法依赖时空网络来学习空间相关性，往往导致对全城所有区域的时空相关性过于泛化的考虑。实际上，一个区域未来的条件很大程度上受到与其具有有用相关性的区域的影响，而不是城市中的每个区域。例如，工作区域的交通由于日常通勤与住宅区本质上是相关的，但与农业区域的相关性有限。因此，我们不是广泛地捕捉所有时空相关性，而是强调识别和利用区域之间的有用相关性来提升全城时空预测。

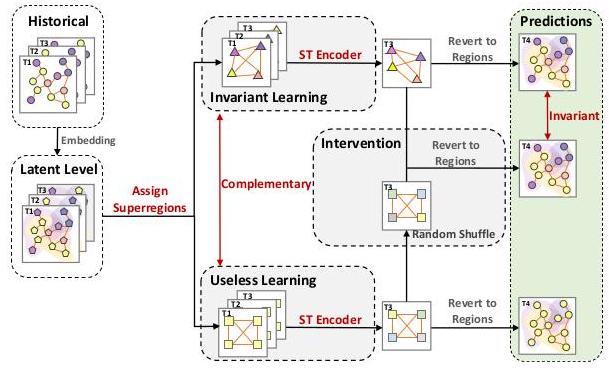


Figure 2: Insight of the network.

图2：网络的洞察。

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 , 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 ; taxi flows are concentrated in downtown districts and sparse in other boroughs . Previous works prioritize regions with large event numbers, often employing region-wise losses, such as MSE, RMSE, to optimize neural networks. Some of them 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.

挑战二：如何限制预测以符合实际全城分布？城市规划分区和功能的特征使得全城时空数据具有独特的分布 ，如图1（b-c）所示。例如，大多数交通事件（如事故）发生在城市区域，它们很少在农村区域发生 ；出租车流量集中在市中心区域，而在其他地区较为稀疏 。先前的工作 通常优先考虑事件数量大的区域，经常使用区域性的损失函数，如均方误差（MSE）、均方根误差（RMSE），来优化神经网络。其中一些研究 甚至通过重新加权的损失策略特别放大高事件区域的影响。然而，这些工作迫使网络计算倾向于事件数量大的区域的误差，忽视了事件数量少的区域。因此，它们可能导致网络生成的预测与实际全城分布有较大偏差。

In this paper, we propose a Causal Attention Network (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:

在本文中，我们提出了一个城市级时空预测的因果注意力网络（CityCAN）。考虑到有用的相关性受到城市规划的影响，我们认为这些相关性是随时间不变的区域间的稳定空间相关性。因此，为了应对挑战I，CityCAN采用了一个因果框架（如图2所示）来学习有用的相关性，同时忽略其补充相关性（即无用的相关性）。在CityCAN中，具有不变/无用的相关性的区域被分配到有用/无用的超区域，用于不变/无用的学习分支，并具有两个补充的超区域矩阵。然后，通过推动不变学习分支和干预分支的预测保持不变，可以识别出有用的相关性，而不管从无用学习分支中学到的无用表示如何变化。为了增强这些分支内的时空建模，我们提出了一个新颖的全局局部注意力编码器（GLAE）作为时空编码器，通过局部和全局的时空注意力联合编码空间和时间依赖性。为了解决挑战II，我们设计了一个城市级损失，该损失从全局角度（即城市层面）惩罚网络。具体来说，它通过考虑城市中的所有区域，约束空间维度上的预测与真实空间分布紧密对齐。换句话说，它衡量了预测与所有区域未来条件之间的分布相似性。总的来说，我们将我们的贡献概括如下：

* We propose CityCAN, a causal attention network for citywide ST forecasting, which leverages causal theory to uncover useful spatial correlations over time.
* 我们提出了CityCAN，一个用于城市级时空预测的因果注意力网络，它利用因果理论来揭示随时间变化的有用空间相关性。
* We introduce a Global Local-Attention Encoder (GLAE) for better spatio-temporal correlation modeling.
* 我们引入了全局局部注意力编码器（GLAE），以更好地进行时空相关性建模。
* We design a citywide loss, which constrains the prediction distribution, leading to improved citywide ST forecasting.
* 我们设计了一个城市级损失，它约束了预测分布，从而提高了城市级时空预测的准确性。
* Experiments show CityCAN outperforms state-of-the-art methods on four datasets in three practical applications.
* 实验表明，CityCAN在三个实际应用的四个数据集上超过了现有最佳方法。

# 2 PRELIMINARIES

# 2 基本概念

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

定义1（区域）：感兴趣的区域，即城市，根据其经纬度 [47] 被划分为 个区域。这些区域可以是规则或不规则的形状。

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 . Given a time interval , where is the dimension of the traffic features.

定义2（交通条件与交通特征）：交通条件是交通相关的条件，例如交通事故数据的风险等级、人群流动数据的流入/流出以及人群密度数据的计数。这些交通条件的特征是交通特征 。给定一个时间间隔 ，其中 是交通特征的维度。

Problem Statement Given observed traffic features with time intervals , spatial adjacency matrix of regions , the goal is to generate interested traffic features for all regions for next time intervals, i.e., . Note that the observed traffic features may have more information than the interested traffic features, i.e., .

问题陈述 给定 时间间隔 的观测交通特征、区域的空间邻接矩阵 ，目标是生成下一个 时间间隔内所有区域感兴趣的交通特征，即 。需要注意的是，观测到的交通特征可能比感兴趣的交通特征包含更多信息，即 。

# 3 CITYCAN

# 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).

在这一节中，我们展示了我们的CityCAN，如图3所示，它采用两种策略来解决城市范围的时空预测问题：一种因果框架来识别有用的时空相关性（第3.1节）和一个城市范围的损失函数来约束预测分布（第3.2节）。

# 3.1 Causal Learning for Citywide Forecasting

# 3.1 城市范围预测的因果学习

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.

由于城市分区和功能性的原因，尽管城市范围数据中的时空相关性在短期内（例如，几天）可能是动态的，但城市区域之间的不变空间相关性（例如，住宅区与学校区域之间的相关性）随时间存在。我们将这些不变的相关性视为有用的相关性。为了识别这些相关性，受到分类任务 [39, 44] 的启发，这些任务采用因果理论来区分相关和不相关的特征，我们从因果的角度观察城市范围的时空预测，并为这个回归任务提出了一种因果学习策略。

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).

3.1.1 因果学习策略。为了识别有用的（即不变的）相关性，如图3(a)所示的CityCAN，采用了一个因果框架，该框架包括输入嵌入层、不变学习分支（ILB）、无用学习分支（ULB）和干预模块（IM）。

Input Embedding transforms raw inputs into the latent level for superregions assignment. We first embed the traffic features into high-level representations through a 2D convolution with kernel size . To inject the space-time location for each region, we extend the positional embedding [71] to ST positional embeddings. Specifically, for a region at time (denoted as ), we define its absolute space-time position as , where the index of and starts from 0 . By adding the representations and ST positional embeddings, we obtain final features for ILB and ULB to assign useful and useless superregions, where is the feature dimension.

输入嵌入层将原始输入转换到潜在层面以进行超区域分配。我们首先通过2D卷积（卷积核大小为 ）将交通特征 嵌入到高级表示中。为了为每个区域注入空间时间位置，我们将位置嵌入 [71] 扩展到ST位置嵌入。具体来说，对于在时间 （表示为 ）的区域 ，我们定义其绝对空间时间位置为 ，其中 和 的索引从0开始。通过将表示和ST位置嵌入相加，我们得到最终特征 ，ILB和ULB利用这些特征分配有用和无用的超区域，其中 是特征维度。

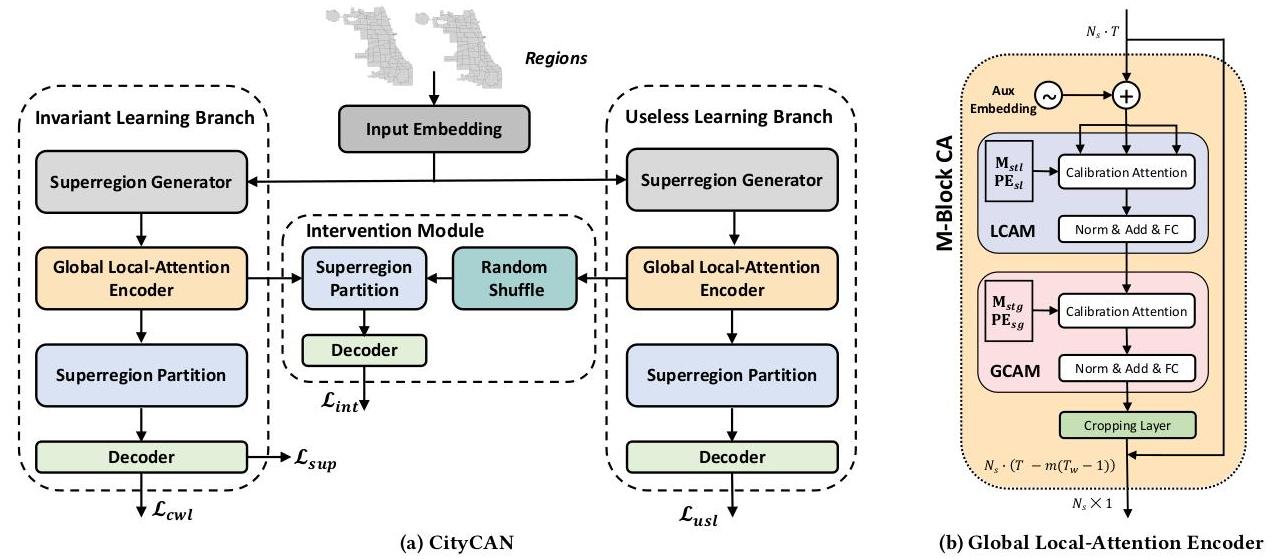


Figure 3: (a) Overview of CityCAN, where invariant and useless correlations among regions are disentangled via the causal learning strategy. (b) Global Local-Attention Encoder (GLAE) learns useful and useless ST features based on invariant and useless correlations. It has the same architecture but different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

图3：(a) CityCAN的概述，其中通过因果学习策略分离了区域之间的不变和无用相关性。（b）全局局部注意力编码器（GLAE）根据不变和无用相关性学习有用的ST特征。它在不变和无用学习分支中具有相同的架构，但参数不同。

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 Super-region Partition, and a Decoder.

不变和无用学习分支（ILB & ULB）分别学习区域之间的不变和无用相关性。它们共享相同的架构，包括超区域生成器、全局局部注意力编码器、超区域划分和译码器。

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 and useless superregion matrix . 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:

超区域生成器将彼此之间存在相关性的区域组合成一个超区域。为了识别有用的相关性并过滤掉无用的相关性，我们引入了两个可学习的超区域矩阵，即有用的超区域矩阵 和无用的超区域矩阵 。它们是从训练数据中观察到的区域之间的相关性派生出来的。为了将相关的/不相关的区域分组到有用的/无用的超区域，我们将这些矩阵应用于原始区域及其相应的相邻关系：

where is the learned adjacent metrics of superregions, is total number of superregions in space dimension, denotes the region reduction parameter, and is the transposition operation. We ensure the useful and useless correlations are complementary to each other, and thus let superregion metrics satisfy , where 1 is the all-one matrix.

其中 是学习到的超区域的相邻度量， 是空间维度中的超区域总数， 表示区域缩减参数， 是转置操作。我们确保有用的和无用的相关性是互补的，因此让超区域度量满足 ，其中 1 是全一矩阵。

Thus, we can obtain useful/useless superregions with their corresponding useful/useless features and adjacent relationships in ILB/ULB. In ILB, the GLAE (our ST Encoder) takes and as inputs, while in the ULB, it uses and 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 for useful superregions and 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 :

因此，我们可以在 ILB/ULB 中获得 个有用的/无用的超区域及其相应的有用/无用特征 和相邻关系 。在 ILB 中，GLAE（我们的 ST 编码器）将 和 作为输入，而在 ULB 中，它使用 和 作为输入。GLAE 旨在捕捉 ST 相关性，无论是在 ILB 中有用的超区域内部还是在 ULB 中的无用区域。它为有用的超区域生成 ST 表示 ，为无用的超区域生成 （更多细节在第 3.1.2 节）。这些表示可以通过使用超区域矩阵 的超区域划分轻松映射回它们原始的区域：

where . Then, given useful ST representation and useless ST representation of original regions, we use the fully-connected layers as the Decoder in ILB and ULB to generate the predictions and useless predictions , where is a task-specific dimension of traffic features. Note that since GLAE is an attention-based encoder, it reduces complexity given the region reduction parameter .

其中 。然后，给定原始区域的有用ST表示 和无用ST表示 ，我们使用全连接层作为ILB和ULB中的解码器来生成预测和无用预测 ，其中 是交通特征的任务特定维度。注意，由于GLAE是一个基于注意力的编码器，它通过区域减少参数 减少了 复杂性。

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 to generate the intervened predictions via the Superregion Partition with and the Decoder. Then, we encourage the invariance between the intervened predictions and the predictions obtained from the ILB to mitigate the impact of useless features through an intervention loss :

干预模块（IM）旨在通过在潜在层面上提供隐式干预来消除无用表示的影响。受到 [44] 的启发，我们首先使用随机洗牌操作生成干预，该操作随机地从不同的无用超区域中收集无用表示。然后，将这些随机干预与有用表示 连接，通过超区域划分 和解码器生成干预后的预测 。接着，我们通过干预损失 促进干预后的预测 与ILB获得的预测 之间的不变性，以减轻无用特征的影响：

where denotes the random shuffle operation, 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.

其中 表示随机洗牌操作， 代表超区域划分和解码器中的操作， 指的是连接函数。为此，CityCAN可以完全利用有用的相关性，通过忽略无用相关性的影响。

Losses for Causal Learning Except for the intervention loss , we also introduce supervised loss and useless loss to disentangle the useful features and useless features for boundless traffic condition values. Supervised loss estimates predictions generated from useful representations in ILB:

因果学习的损失 除了干预损失 之外，我们还引入了监督损失和无用损失，以分离有用特征和无用特征，以便于处理无限交通条件值。监督损失 估计由ILB中的有用表示生成的预测：

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

与使用均匀分类损失消除不相关模式影响的分类工作 [44] 不同，我们为回归任务设计了一个无用损失 ，通过将其值减小到零来推动无用表示变得不必要：

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

然后，因果学习策略的总损失为：

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

其中 是超参数。为此，CityCAN 利用因果理论中的干预策略，指导因果框架识别区域间的有用相关性。

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 to address short- and long-term temporal dependencies, GLAE extends the temporal attention to the spatio-temporal attention for citywide ST forecasting. A GLAE has 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.

3.1.2 全局局部注意力编码器（GLAE）。如3.1.1节所述，我们提出了GLAE，如图3（b）所示，以捕捉ILB和ULB中的ST相关性。由于车辆在城市的行驶速度可以不同，或快或慢，因此建模局部和全局ST相关性至关重要。受到最近研究使用时间注意力 处理短期和长期时间依赖性的启发，GLAE将时间注意力 扩展到城市范围ST预测的时空注意力。GLAE具有 校准注意力（CA）模块，包括辅助特征嵌入、局部CA模块（LCAM）、全局CA模块（GCAM）和裁剪层。由于ILB和ULB中GLAE的结构相同，我们在后续章节中省略了具体的分支名称。

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 , the ST positional index of its location is , where and is the index of the space location and time location, respectively. We can obtain the ST Pos embedding 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 . Then, we obtain the Aux embedding by concatenating the ST Pos embedding and Ext embedding , i.e., . After that, we obtain inputs for subsequent ST correlation modeling by adding the Aux embedding to ST representations .

辅助特征（Aux）嵌入提供了ST位置信息和外部因素信息。它包括ST位置（Pos）嵌入和外部因素（Ext）嵌入。为了增强位置上下文，我们引入ST Pos嵌入来编码时空位置。它将标准位置嵌入[71]扩展为ST格式。对于一个超区域 ，其位置的ST位置索引是 ，其中 和 分别是空间位置和时间位置的索引。我们可以通过对ST位置索引应用可学习的位置嵌入[46]来获得ST Pos嵌入 。为了将外部因素注入潜在特征，我们使用可学习的嵌入层来编码外部因素并生成Ext嵌入 。然后，我们通过连接ST Pos嵌入 和Ext嵌入 ，即 ，来获得Aux嵌入。之后，我们通过将Aux嵌入 添加到ST表示 中，来获得后续ST相关性建模的输入。

CA Module for Local & Global ST Learning The conventional attentions 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:

用于局部和全局ST学习的CA模块 传统注意力 不能直接应用于ST数据，因为它们主要关注时间维度，忽略了ST关系。城市范围的ST数据有两个关键ST关系：（1）未来的交通条件不能影响过去的情况；（2）空间上连接的超区域相互影响更大。为了将这些关系纳入注意力操作，我们提出了一个校准注意力模块（CAM），其核心是校准注意力（CA）操作。CA操作通过两个组件来根据城市范围的ST关系校准注意力：

* ST influential mask 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 is not a triangular matrix as the superregions are arranged by space-time location.
* ST影响掩码 通过将代表未来时间间隔影响的注意力设置为0来防止未来信息泄露 [46]。然而，与先前工作中的掩码不同 ，它不是一个三角矩阵，因为超区域是按空间时间位置排列的。
* Spatial bias enhances spatial relationships by setting temporal influence to zero and repeating spatial influence with superregions’ spatial relationships .
* 空间偏差 通过将时间影响设置为0并重复超区域的空间关系中的空间影响 来增强空间关系。

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

然后，我们将传统的注意力操作修改为校准注意力（CA）操作：

where , and . Note that reshaping and broadcasting are needed to retain the ST positional index of superregions. In the CAM, CA operations are performed to attend different ST patterns. Then, the output of the CAM is the aggregated ST representations , obtained by applying a layer normalization, a residual connection, and a fully connected feed-forward network on the concatenation of the CA operations.

其中 ，以及 。注意，需要重塑和广播以保留超区域的ST位置索引。在CAM中， CA操作被用来关注不同的ST模式。然后，通过在 CA操作的串联上应用层归一化、残差连接和全连接的前馈网络，得到CAM的输出，即聚合的ST表示 。

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):

CAM由于其基于注意力的设计，能够捕捉局部和全局ST特征。为了更好地学习局部和全局ST特征，我们以两种不同的方式使用CAM：局部CAM（LCAM）和全局CAM（GCAM）：

Local CAM (LCAM) captures local ST representations within a sliding window of size . The total number of ST superregions in each window is . We apply CAM on these supperregions with their components (see calculations above). Given temporal intervals at -th block, there are sliding windows, resulting in superregions. The index of 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., , where is the total number of superregions in -th block, and is:

局部 CAM（LCAM）在大小为 的滑动窗口内捕获局部 ST 表征。每个窗口中的 ST 超区域总数为 。我们在这些超区域及其组成部分 上应用 CAM（参见上述计算）。给定 个时间间隔在 -th 块上，有 个滑动窗口，导致 个超区域。 的索引从 1 开始。由于每个超区域聚集了来自所有其他超区域的 ST 特征，我们让 LCAM 只输出每个滑动窗口的最后一个时间间隔的特征，即 ，其中 是 -th 块中的超区域总数， 是：

Global CAM (GCAM) learns the global ST features from all super-regions across time and space. Similar to LCAM, we apply CAM on all ST superregions, i.e., superregions, with their corresponding calibration components , to obtain the final output of GCAM .

全局 CAM（GCAM）从所有超区域跨时间和空间学习全局 ST 特征。与 LCAM 类似，我们在所有 ST 超区域上应用 CAM，即 个超区域，以及它们对应的校准组成部分 ，以获得 GCAM 的最终输出 。

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 because each superregion has aggregated ST information from all other superregions in GCAM. Thus, at the last block , we only use the superregions at the last temporal interval , i.e., . Then, total number of superregions is updated to:

裁剪层从最远超区域中移除冗余特征，因为交通条件主要受最近时间间隔的影响。冗余信息存在于 中，因为每个超区域在 GCAM 中已经聚集了来自所有其他超区域的 ST 信息。因此，在最后一个块 上，我们只使用最后一个时间间隔的超区域 ，即 。然后，超区域的总数 更新为：

# 3.2 Citywide Loss

# 3.2 城市范围损失

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.

城市内的区域，受到城市规划分区和功能性的影响，可以分为以下几类：（1）重要区域，特点是频繁发生事件，可能需要额外的人力干预（例如，在预测到事故发生时进行交通管制或为预测需求高的区域预分配出租车）。（2）次要区域，通常事件数量较小或为零，不需要特定的人力干预。重要区域和次要区域在城市的分布是不均匀的。为了确保有效的干预且不浪费资源，网络应能准确预测城市所有区域的目标特征。然而，因果损失（式6）强调区域错误，可能导致预测结果与城市的实际空间分布不一致。为了解决这个问题，我们引入了一个辅助损失，名为全局损失，用以规范预测值和标签之间的分布。同时，考虑到重要区域在需要大量人力干预的应用中具有特别重要的地位，我们首先引入了一个校准先验，以增加重要区域的权重。

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 by summarizing the interested conditions features of each region over observed samples, i.e., training samples, and obtain the calibration prior based on the region prior :

校准先验利用了城市范围内的领域知识，即相似的空间分布随时间存在。这种知识存在是因为交通受到城市地理和语义的影响。因此，我们可以通过一个区域先验 识别重要区域，通过总结每个区域在观测样本（即训练样本）中的感兴趣条件特征，并基于区域先验 获得校准先验 ：

where is the total number of training samples, is the index of spatial region, refers to targeted traffic condition features, e.g., the features of traffic accident risk, taxi flow, crowd density, and is the calibration parameter that controls the selection of the most important significance regions.

其中 是训练样本的总数， 是空间区域的索引， 指的是目标交通条件特征，例如交通事故风险特征、出租车流量、人群密度， 是校准参数，用于控制选择最重要的显著性区域。

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:

城市范围损失校准先验使网络能够生成可以反映真实城市分布的预测分布，同时更多地惩罚重要区域中的错误。我们基于重新加权的余弦相似度计算城市范围损失：

where 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.

其中 是一个超参数，用于避免除以零。重新加权的余弦相似度集体应用于每个时间间隔的所有区域，因此可以约束交通状况的空间分布。在训练过程中，它还通过对城市中各个区域的重要性进行重新加权，从而对每个区域提供适当的关注。请注意，校准先验同时应用于预测和标签，从而保持分布。

# 3.3 Losses for CityCAN

# 3.3 城市CAN的损失函数

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

城市CAN的最终损失包含两部分，即公式6中的因果损失和公式11中的城市范围损失（CWL）：

where 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.

其中 是超参数。为此，CityCAN可以同时考虑区域性和城市范围的错误，因此确保了在整个城市范围内的高预测性能。

# 4 EXPERIMENTS

# 4 实验部分

# 4.1 Experimental Settings

# 4.1 实验设置

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 contains irregular regions, better representing natural city divisions. Dataset details are in Table 1.

4.1.1 数据集。我们在四个真实世界的数据集上评估CityCAN，即NYC13 [47]、BikeNYC [67]、Chicago21和Chicago22。NYC13和BikeNYC是基于网格的数据集，具有规则的区域，而Chicago21和Chicago22数据集 包含不规则区域，更好地代表自然城市划分。数据集的详细信息在表1中。

Table 1: The statistic of datasets.

表1：数据集的统计数据。

| Dataset | NYC13 | BikeNYC | Chicago21 | Chicago22 |
| --- | --- | --- | --- | --- |
| Time Span | 01/01/2013 - | 04/01/2014 - | 01/01/2021- | 01/01/2022 - |
| (mm/dd/yyyy) | 12/31/2013 | 09/30/2014 | 12/31/2021 | 12/31/2022 |
| Time Interval Region Size | 1 hour | 1 hour (16, 8) | 30 minutes 77 | 30 minutes 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:

4.1.2 任务。我们在三项任务上开展实验，包括交通事故风险预测、人群流动预测和人群密度预测：

Traffic accident risk forecasting task: we follow the existing works , 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 given historical observations of 6 -time intervals .

交通事故风险预测任务：我们遵循现有研究 ，不仅要预测交通事故的发生，还要估算风险值。风险值应反映该区域交通事故的频率和严重程度，因此将其定义为区域内每起交通事故严重程度的总和。在实验中，我们根据前6个时间间隔的历史观测数据 ，预测下一个时间间隔 的交通事故风险情况。

Crowd flow forecasting task and crowd density task: we follow existing works 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 given historical observations of 6-time intervals .

人群流动预测任务和人群密度任务：我们遵循现有研究 ，分别预测城市中所有区域的人群流入/流出和人群密度值。在实验中，我们根据前6个时间间隔的历史观测数据 ，预测接下来的6个时间间隔的人群流动情况和人群密度情况 。

4.1.3 Evaluation Metrics. We follow the previous studies 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 of the model to indicate regions with risk, where F1@K denotes the F1 score for top regions with high accident values.

4.1.3 评价指标。我们遵循以前的研究 ，使用两个指标来评估我们的模型：平均绝对误差（MAE）和均方根误差（RMSE）。此外，为了更全面地评估交通事故风险预测，我们使用F1得分、F1@20、F1@30来展示模型指示风险区域的能务，其中F1@K表示前 个高风险区域的F1得分。

https://data.cityofchicago.org/

https://data.cityofchicago.org/

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 to 4, number of blocks to 4, feature dimension to 128, number of multi-heads to 4 . We balance the region-wise and city-wise losses and set to , and 0.5, respectively. The calibration parameter 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.

4.1.4 实施细节。我们的模型在单个 GTX 2080 Ti 上使用 Adam 优化器进行训练，学习率为 0.001。我们将区域缩减参数 设为 4，模块数量 设为 4，特征维度 设为 128，多头数量 设为 4。我们平衡了区域级和城市级的损失，并分别将 设为 和 0.5。校准参数 是特定于数据集的。我们采用提前停止策略，所有实验的最大迭代次数为 150 个周期。对于数据划分，我们使用最后 8 周作为测试集，前 4 周作为验证集，其余数据作为训练集。

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

4.1.5 基线。为了充分展示 City-CAN 在不同任务中的有效性，我们采用了以下针对每个任务特别设计的基线：

* Classical methods: HA [4] averages the historical traffic conditions of the same time slot given the past observed segments.
* 经典方法：HA [4] 对过去观察到的路段在相同时间段的交通条件进行平均。
* 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.
* 交通事故风险预测：我们选择了三种流行方法，即 SDAE [6]、SDCAE [5] 和 GSNet [47]，以及两种通用时空模型，即 STGCN [63] 和 GWNET [58]，作为基线。
* 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].
* 人群流动和人群密度预测：我们选择了六个强大的模型进行比较，包括 STGCN [63]、DCRNN [27]、GWNET [58]、AGCRN [2]、MTGNN [57] 和 GMSDR [31]。

# 4.2 Experimental Results & Analysis

# 4.2 实验结果与分析

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. 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.

4.2.1 交通事故风险预测。表2展示了基线模型和我们的模型在两个数据集上对交通事故风险的预测结果。我们的模型在所有数据集上，就风险值的准确性和事故指示的F1值而言，始终优于基线模型。需要注意的是，最近的研究工作，例如SDCAE和GSNet，无法适应Chicago21数据集。这是因为这些为规则区域设计的模型，使用了CNN进行空间捕捉。然而，Chicago21数据集反映了一个城市自然的社区划分，包含了不规则区域。这些区域具有非欧几里得结构，无法使用CNN建模。我们的模型平均提高了2.07倍的全市F1值，证明了其优越的地区识别能力，无论该地区是否发生事故。（2）针对交通事故风险预测设计的基线模型（即SDAE、SDCAE和GSNet）比一般的ST预测模型（即STGCN和GWNET）表现得更好。这是因为事故是罕见事件，而一般的ST模型专注于ST建模，未能考虑到这项任务中固有的稀疏数据。（3）令人惊讶的是，HA在全市F1得分上超过了所有基线模型。我们推测，深度模型专注于重要区域，忽略了次要区域，因此无法识别次要区域，导致全市F1得分较低。另一方面，HA只考虑每个区域的历史观测数据，避免了这个问题。我们的模型采用全市损失函数，考虑预测分布，并适当关注每个区域，因此取得了最佳性能。（3）我们的模型实现了最低的MAE和RMSE误差，表明它能生成更准确的风险值，因为它基于真正影响未来状况的有用相关性生成预测。

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

表2：在NYC13和Chicago21数据集上的交通事故预测模型比较，其中 - 表示模型无法应用于该数据集。

| Dataset | Method | MAE | RMSE |  | F1@30↑ | F1@20↑ |
| --- | --- | --- | --- | --- | --- | --- |
| NYC13 | HA | 0.05 | 0.28 | 11.10% | 10.98% | 11.02% |
| SDAE | 0.07 | 0.48 |  | 32.38% | 37.45% |
| SDCAE | 0.10 | 0.70 |  | 52.88% | 60.17% |
| GSNet | 0.04 | 0.25 | 7.19% | 24.45% | 26.09% |
| STGCN | 0.05 | 0.28 | 3.97% |  | 4.36% |
| GWNET | 0.04 | 0.27 | 3.01% | 3.11% |  |
| 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

从结果中，我们可以观察到：（1）我们的模型在事故指示上比基线模型大幅超出。具体来说，它

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

表3：在BikeNYC数据集和Chicago21数据集上的客流预测模型比较，以及在Chicago22数据集上的 crowd density 预测。

| 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.

4.2.2 人群流动预测与人群密度预测。

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 and RMSE error by on average over three datasets. It demonstrates that our proposed model is a general model which can achieve better performance on various citywide tasks.

表3展示了基线模型和我们的模型在BikeNYC和Chicago21数据集上的客流预测，以及在Chicago22数据集上的人群密度预测的结果。与事故风险预测任务相比，这两项任务的数据并不稀疏。结果显示，我们的模型在所有指标上始终优于现有方法。具体来说，它在三个数据集上平均减少了MAE误差 和RMSE误差 。这表明我们提出的模型是一种通用模型，能够在各种城市范围内的任务上实现更好的性能。

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

表4：CityCAN在Chicago22数据集上进行人群密度预测的消融研究。

| 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

# 4.3 消融研究

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

表4详细说明了CityCAN中每个组件的有效性。 缺乏因果框架，它采用了一个单一的不变学习模块。w/o UIL 是没有无用损失（公式5）和干预损失（公式3）的模型。w/o UL 排除了无用损失。w/o IL 省略了干预损失。w/o CWL 是没有城市范围损失（公式11）的模型。w/o CP 在城市范围损失内没有校准先验。w/o CL 移除了裁剪层。w/o LCAM 和 w/o GCAM 分别是没有LCAM和GCAM模块的模型。

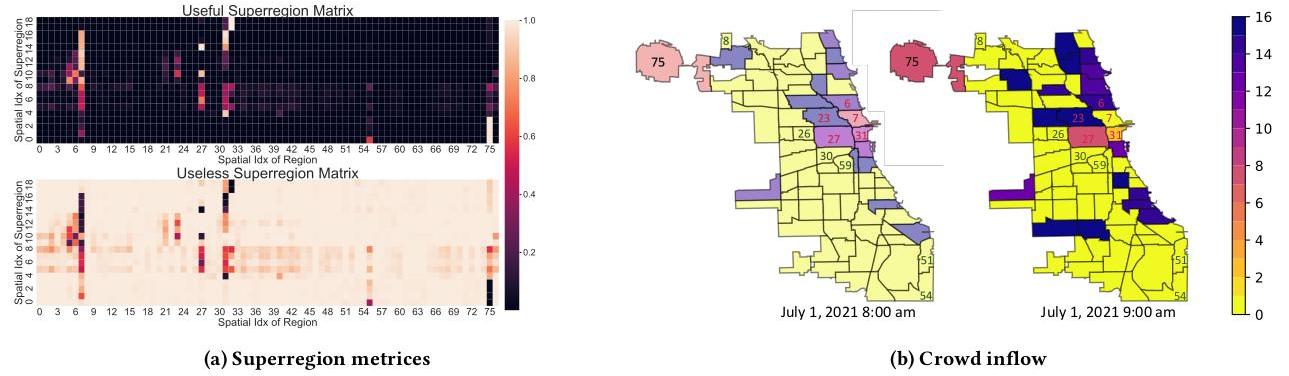


Figure 4: (a) A visualization of superregion metrices for crowd flow forecasting on Chicago21 dataset. (b) A visualization of crowd inflow on Chicago21 dataset, where the idx (index) in each region refers to its spatial index.

图4：(a) 芝加哥21数据集上人群流动预测的超区域度量可视化。(b) 芝加哥21数据集上人群流入的可视化，其中每个区域的idx（索引）指的是其空间索引。

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 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 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.

表4揭示了：(1) 因果学习策略提高了性能，验证了识别有用相关性的有效性。(2) 排除无用的损失或干预损失损害了模型性能，表明无用的相关性确实存在，并且误导了网络。这两种损失必须共同作用以实现因果学习，无用的损失确保了无用特征为零影响，而干预损失保证了添加无用特征后的结果不变。(3) 省略全市损失降低了性能，证明了考虑所有城市区域的重要性。(4) 中较高的RMSE表明，校准先验可以提供有用的领域知识，以提高条件值高区域的性能。尽管应用校准先验倾向于更多地关注RMSE，导致MAE略有增加，但在条件值高的场景中，如交通事故风险，它特别有用。其影响可以通过将校准参数设置为1来轻松移除。(5) CL的较差性能突出了移除冗余信息可以使模型专注于最重要的特征。(6) 不带LCAM和不带GCAM的表现较差，证明了捕捉局部和全局ST相关性对于全市ST预测是必要的。

# 4.4 Visualization

# 4.4 可视化

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

4.4.1 超区域矩阵。图4展示了两个超区域矩阵，即有用的超区域矩阵和无用的超区域矩阵，用于在Chicago21数据集上进行人群流动预测。从图4中，我们可以观察到：（1）这两个矩阵互为补充，成功地从无用的关联中区分出了有用的关联。（2）由于具有有用关联的区域被分组到同一个超区域中，因此发现了有用的关联。例如，像区域7和区域31这样的中央商务区（CBD）区域，以及像区域27这样的住宅区域，被分配到同一个超区域（例如，超区域14）。这种分组表明了它们的关联，并与已知的城市研究洞察相符。（3）某些乡村区域，例如区域8和区域51，与其他区域的相关性较少，因此在无用的超区域矩阵中权重较高。（4）我们的模型允许每个区域根据其与不同区域的相关性被分配到多个超区域。例如，区域7也被分配到超区域9，因为它还与区域5和区域6有相关性。一个区域对不同的超区域的权重根据其在不同关联中的重要性而变化。

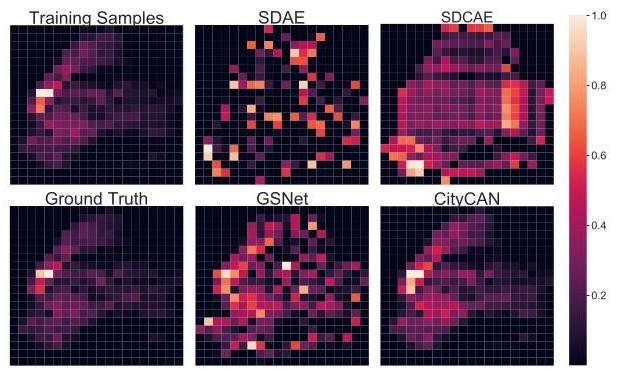


Figure 5: A visualization of the citywide spatial distribution of traffic accidents in the NYC13 dataset, along with the forecasting results of various methods.

图5：纽约市NYC13数据集中交通事故全市空间分布的可视化，以及各种方法的预测结果。

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.

51号区域与其他区域的相关性较少，因此在无用的超区域矩阵中权重较高。（4）我们的模型允许每个区域根据其与不同区域的相关性被分配到多个超区域。例如，区域7还被分配到超区域9，因为它还与区域5和区域6有相关性。一个区域对不同的超区域的权重根据其在不同关联中的重要性而变化。

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

4.4.2 城市范围分布。图5显示了纽约市数据集（NYC13）上城市范围数据的分布情况，这是通过对时间维度的交通事故风险值进行平均，并对空间维度的交通事故风险值进行归一化得到的。训练样本的可视化来源于训练集，而真实情况代表了测试集中预测的真实情况的可视化。这两种可视化具有相似的分布，验证了我们的假设，即城市范围分布的训练集和测试集之间没有显著变化。SDAE、SDCAE和GSNet模型产生的预测可视化说明了这些模型在生成与实际城市范围分布一致的良好预测方面的局限性。这是因为它们采用了基于区域的损失函数，忽略了城市范围的分布。特别是，它们无法识别平凡区域，这些区域在时间上持续呈现零值，可能会导致资源分配低效。CityCAN在预测时同时考虑了基于区域和城市范围的误差，因此可以

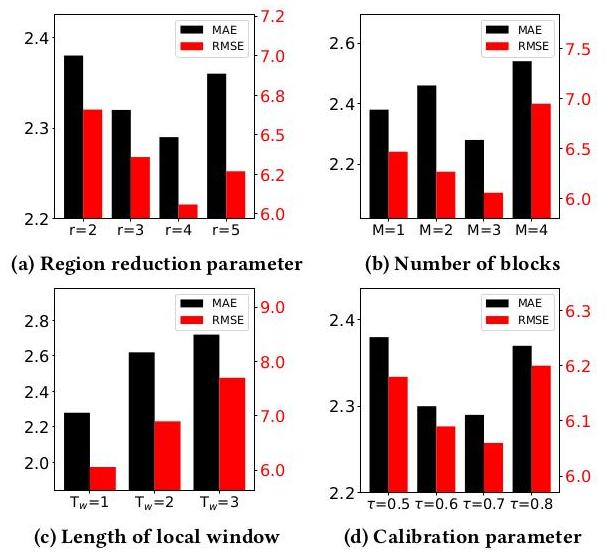


Figure 6: Effects of hyperparameters on Chicago22 dataset in terms of MAE and RMSE.

图6：在MAE和RMSE方面，超参数对Chicago22数据集的影响。

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

成功识别重要区域和平凡区域，并对城市所有区域的预测结果都很好。

# 4.5 Effects of Hyperparameters

# 4.5 超参数的影响

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 . 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 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 , particularly in terms of RMSE, because it gives suitable weights to the significant regions that have higher values.

在图6中，我们研究了超参数在CityCAN中对芝加哥22进行人群密度预测的影响。从结果中我们观察到：（1）当区域缩减参数 时，CityCAN实现了最低的均方根误差（RMSE）。这是因为更高的缩减率导致超级区域数量减少，这使得网络能够从原始相关区域获取概括性特征并消除一些冗余特征。然而，如果缩减率过高，它可能包含许多无用的相关性，这将对性能产生负面影响。（2）当包含 块时，CityCAN实现最高性能，因为减少/增加块的数量可能导致欠拟合/过拟合问题。（3）增加局部窗口长度将对性能产生负面影响，因为在短期预测中考虑每个时间间隔的时间信息是很重要的。（4）当 时，CityCAN表现最佳，特别是在均方根误差（RMSE）方面，因为它为具有较高值的显著区域赋予了适当的权重。

# 5 RELATED WORK

# 5 相关工作

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 , traffic flow prediction , traffic speed prediction , taxi demand prediction , etc. They have achieved superior performance over traditional statistical models, like k-nearest neighbor [35] and ARIMA thanks to their ability to model complex nonlinear ST correlations. More recent works 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 , 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 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.

城市范围的空间时间预测是智能交通系统（ITS）的一项关键任务，多年来一直备受关注。近期的研究探讨了用于各种城市范围任务的空间时间（ST）网络，例如交通事故预测 ，交通流量预测 ，交通速度预测 ，出租车需求预测 等。由于它们能够建模复杂非线性ST相关性，这些网络在传统统计模型，如k-最近邻 [35] 和ARIMA 的基础上实现了优越的性能。更近的研究 表明，联合学习空间和时间依赖性可以增强预测性能。然而，它们在同时考虑不规则区域之间的全局ST相关性时仍面临挑战。同时，基于注意力的模型 [56, 71] 在时间预测上学习全局动态依赖性方面取得了成功。但是，它们专注于长期多变量时间序列和有效的注意力机制 ，忽略了空间相关性和领域知识，因此不能直接应用于城市范围ST预测。此外，在城市范围预测中，城市范围的分布相对未被充分探索。尽管有些研究 已经研究了在城市的稀疏分布中的零膨胀数据。但它们关注的是区域优化，这导致产生了与城市范围分布不符的偏斜预测。在这项工作中，我们提出了一种新颖的基于注意力的ST编码器，该编码器在因果关系框架中融入了城市范围的领域知识，并引入了一种城市范围损失，以约束预测分布，实现更好的ST建模。

Causal Learning 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].

因果学习 赋予深度学习模型消除伪相关性的能力，从而在各种任务中提高性能。例如，CONTA [66] 通过在图像分割任务中的后门调整，移除图像像素与标签之间的非因果关联。Liu等人[34]通过解耦物理定律、风格混淆因素和非因果特征，学习运动表示的因果不变性，以实现更好的运动预测。CAL [44] 通过在表示层面上应用因果干预，提升图分类性能。STNSCM [10] 分析输入数据与上下文条件之间的因果关系。与它们不同，我们提出了一个因果注意力网络，用于移除城市范围回归中ST数据中存在的无用相关性。还有一些关于ST数据因果学习的研究[59, 74]。

# 6 CONCLUSION AND FUTURE WORK

# 6 结论与未来工作

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.

在本文中，我们提出了CityCAN，一种用于城市范围ST预测的新型网络。利用因果理论，我们为城市范围ST预测设计了一个因果框架，该框架在潜在层面上应用隐式干预，使CityCAN能够学习有用的区域相关性。为了共同捕捉规则和不规则区域的ST相关性，我们还在CityCAN中引入了一个全局-局部注意力编码器。它通过校准的注意力机制捕捉局部和全局ST相关性，以实现更好的ST建模。然后，我们提出了一个城市范围损失函数，该函数考虑了预测条件与现实条件之间的城市范围分布，使CityCAN能够一次性准确预测城市所有区域的目标特征。广泛的实验结果和分析验证了CityCAN的有效性。CityCAN不仅限于城市范围ST预测。在未来，我们将在其他ST任务上评估它，例如犯罪预测。我们还将探索不同的架构用于不变学习和无用学习，以降低计算成本。

# 7 ACKNOWLEDGEMENT

# 7 致谢

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.

本研究得到了新加坡国家研究基金会（National Research Foundation, Singapore）在其行业对齐基金 - 预定位（IAF-PP）资助计划下的支持。本材料中表达的任何观点、发现、结论或建议都是作者（们）的观点，并不反映新加坡国家研究基金会的观点。

# REFERENCES

# 参考文献

[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 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.

[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. SpatioTemporal 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.

[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

[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. 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

[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 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 Bench- (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,

[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: Cam-bridgeUniversityPress 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),

[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. Lin-former: 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 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.

[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. 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 rides-ourcing 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

[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

[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.