C^{3}DE: Causal-Aware Collaborative Neural Controlled Differential Equation for Long-Term Urban Crowd Flow Prediction

Yuting Liu[1], Qiang Zhou[1] (✉), Hanzhe Li[1], Chenqi Gong[2], and Jingjing Gu[1]

[1] Nanjing University of Aeronautics and Astronautics, Nanjing, China
{yuting_liu,zhouqnuaacs,lihanzhe,gujingjing}@nuaa.edu.cn
[2] Chongqing University, Chongqing, China
gcq@stu.cqu.edu.cn

Abstract. Long-term urban crowd flow prediction suffers significantly from cumulative sampling errors, due to increased sequence lengths and sampling intervals, which inspired us to leverage Neural Controlled Differential Equations (NCDEs) to mitigate this issue. However, regarding the crucial influence of Points of Interest (POIs) evolution on long-term crowd flow, the multi-timescale asynchronous dynamics between crowd flow and POI distribution, coupled with latent spurious causality, poses challenges to applying NCDEs for long-term urban crowd flow prediction. To this end, we propose Causal-aware Collaborative neural CDE (C^{3}DE) to model the long-term dynamic of crowd flow. Specifically, we introduce a dual-path NCDE as the backbone to effectively capture the asynchronous evolution of collaborative signals across multiple time scales. Then, we design a dynamic correction mechanism with the counterfactual-based causal effect estimator to quantify the causal impact of POIs on crowd flow and minimize the accumulation of spurious correlations. Finally, we leverage a predictor for long-term prediction with the fused collaborative signals of POI and crowd flow. Extensive experiments on three real-world datasets demonstrate the superior performance of C^{3}DE, particularly in cities with notable flow fluctuations.

Keywords: Long-Term Urban Crowd Flow Prediction · Neural Controlled Differential Equation · Counterfactual Inference.

1 Introduction

Urban development is a dynamic process driven by population growth, economic activities, and infrastructure development. As a key indicator of urban operations, urban crowd flow exhibits a significant continuous evolution trend. Exploring its long-term evolution helps reveal urban operating patterns and provides valuable insights for traffic management and sustainable urban development.

Existing researches on long-term prediction [15,32,34] typically employed coarse-grained data with hourly or even longer intervals, in contrast to the high-frequency, minute-level data commonly used. Such coarse-grained data obscures the important urban dynamics and trends, leading to information loss and making it harder for models to capture crowd flow dynamics, resulting to suboptimal prediction performance. Consequently, we introduce a continuous modeling approach to better capture urban dynamics from coarse-grained data, enhancing prediction stability and accuracy. Intuitively, it is essential to consider the evolution of urban structure in urban crowd flow prediction, which is mainly reflected in the changes of POI distribution [19,33]. For example, the construction of a new commercial center may attract higher pedestrian flow, while the renovation of an old residential area may affect the surrounding traffic flow.

However, modeling the impact of POI distribution on crowd flow (also referred to as collaborative signals) from a continuous-time perspective poses the following two challenges: ***i) The multi-timescale asynchronous dynamics of collaborative signals increase the difficulty of modeling spatio-temporal dependencies in urban dynamic systems.*** The evolution of collaborative signals occurs across different time scales, with their dynamic changes unsynchronized. Specifically, changes in low-frequency POI distributions gradually manifest in the high-frequency crowd flow patterns. This cross-scale influence is often reflected in significant crowd flow variations across multiple timestamps, which significantly increases the difficulty of modeling the multi-scale asynchronous dynamic and revealing dynamic patterns in a urban system. ***ii) The accumulation of spurious correlations between collaborative signals complicates the identification of true causal relationships in continuous modeling.*** POI distribution and urban crowd flow often exhibit statistically spurious correlations, which may mislead the model. From a discrete-time perspective, spurious correlations can be easily identified and removed through statistical methods like calculating correlation coefficients or Granger causality tests [7]. However, in continuous-time modeling, where time is treated as a continuous variable and dynamics are learned through differential equations[4], spurious correlations can be amplified during long-term integration. Moreover, minor perturbations in continuous time can significantly impact the overall system, further complicating the accurate identification of true causality.

To this end, we propose a **C**ausal-aware **C**ollaborative Neural **C**ontrolled **D**ifferential **E**quations framework (C^{3}DE) for long-term urban crowd flow prediction. Specifically, we propose a collaborative neural controlled differential equation (NCDE) with a dual-path architecture to capture the dynamic evolution of collaborative signals across different timescales in continuous time. With the continuous-time integration property of NCDE, the asynchronous dynamics of collaborative signals can be effectively modeled. Furthermore, we design a counterfactual-based causal effect estimator to simulate urban dynamics under different POI distribution interventions, enabling a quantitative assessment of each POI category's direct impact on crowd flow. To mitigate the accumulation of spurious correlations among collaborative signals, we incorporate causal effect values into the NCDE framework and introduce a causal effect-based dynamic correction mechanism. By computing causal influences across multiple time steps and feeding them back into POI representations, the mechanism effectively minimizes interference from spurious POIs, alleviates the amplification of spurious correlations, and enhances the model's robustness and reliability in the long-term prediction task.

Overall, our contributions can be summarized as follows:

- To the best of our knowledge, C^{3}DE, is the first to simulate the evolution of collaborative signals and explore the underlying causal mechanisms for long-term urban crowd flow prediction.
- We propose a collaborative NCDE with a dual-path architecture to effectively capture the asynchronous evolution of collaborative signals across multiple timescales.
- We design a counterfactual-based causal effect estimator to quantify the causal impact of POIs on crowd flow and introduce a causal effect-based dynamic correction mechanism to reduce the accumulation of spurious correlations.
- Extensive experiments on three real-world datasets demonstrate that C^{3}DE offers a significant advantage in modeling crowd flow dynamics, particularly in cities with notable flow fluctuations.

2 Related Work

Urban Crowd Flow Prediction. Recently, urban crowd flow prediction [2,21] has become a critical research topic, relying on historical flow data and using Gated Recurrent Unit (GRU) and Graph Neural Networks (GNNs) to learn spatio-temporal features. Traditional spatio-temporal GNNs, like STGCN [31] and STSGCN [26], used predefined graph structures to capture spatial dependencies but often fail to capture the hidden ones. To address this, methods based on adaptive graph structures introduced learnable adaptive adjacency matrices, enabling capturing the dynamics of node relationships [1,25,30]. In addition, considering that urban structure, i.e., POI distribution, significantly affect crowd mobility patterns, some works have incorporated it into flow pattern modeling [20,23]. For example, GeoMAN [20] treated POI as a spatial feature to capture spatial correlations within regions. GSTE-DF [23] utilizes POI data to uncover differences and similarities between regions for inferring origin-destination flows. Although these works achieved some success, they treated POI as a static feature and ignored the dynamics of POI distribution in cities.

Neural Ordinary Differential Equations. [4] first combined neural networks with Ordinary Differential Equations (ODE) and proposed Neural ODEs to model continuous dynamics, which has been widely used in the fields of time series prediction [11,16], continuous dynamic systems [12,13]. [10] proposed tensor-based ODEs to capture spatio-temporal dynamics, overcoming the limitations of graph convolutions in modeling long-range spatial dependencies and semantic connections. [5] designed two types of Neural Controlled Differential Equations to handle temporal and spatial dependencies separately. Additionally, [22] proposed STDDE, which incorporates delayed states into NCDE, allowing it to model time delays in spatio-temporal information propagation.

Counterfactual Inference. The main goal of counterfactual inference is to analyze potential outcomes through hypothetical interventions and answer "What would have happened if the situation had been different?" [3,27]. For example, [18] proposed a counterfactual data augmentation-based causal explanation framework that identifies the true causal factors by constructing counterfactual data. [24] introduced a counterfactual explanation method based on causal intervention, using a causal director to capture causal relationships in the distribution and guide counterfactual generation. In this paper, We address the spurious correlations between collaborative signals from a counterfactual perspective.

3 Preliminary

3.1 Definitions and problem statement

Definition 1 (Urban Network). The urban network is represented as a directed graph $G = (V, X, A)$, where $V = \{V_1, V_2, ..., V_N\}$ denotes N regions in the city. $X \in \mathbb{R}^{t' \times N \times C}$ denotes the urban crowd flow across N regions at t' time steps, where t' is measured in days and C capturing hourly features. $A \in \mathbb{R}^{N \times N}$ is the adjacency matrix, which encodes the relationships between regions.
Definition 2 (POI Distribution). The POI distribution is denoted as $P \in \mathbb{R}^{t'' \times N \times K}$, where t'' is the time steps, measured in months. K represents the number of POI categories, such as restaurants, shops and public facilities.
Problem Statement (Long-Term Urban Crowd Flow prediction). Given the crowd flow for the past T time steps and POI distribution for the past M time steps, our goal is to learn a map function $\mathcal{F}(\cdot)$ that capture the causal evolutionary relationship and predict the urban crowd flow for the next S time steps. It can be formulated as follows:

$$\mathcal{F}^* = arg \min_{\mathcal{F}} \sum_{S} \ell(\mathcal{F}(X_{t'-T+1:t'}, P_{t''-M+1:t''}), X_{t'+1:t'+S}), \tag{1}$$

where $\mathcal{F}^*$ denote the function with the learned optimal parameters, and $\ell(\cdot)$ is the loss function.

In this work, we divide the map function $\mathcal{F}(\cdot)$ into two stages, i.e., a representation part $F(\cdot)$ to model the collaborative causal evolutionary relationship and a predictor $G(\cdot)$ to predict the future crowd flow.

3.2 Neural Differential Equation

Neural ODEs. Neural ODEs [4] extend residual networks into the continuous time domain. Given the input X, neural ODEs define a hidden state $h(t)$ that evolves over time t, as described by the following Riemann integral:

$$h(t) = h(0) + \int_0^t \frac{\mathrm{d}h}{\mathrm{d}t} \mathrm{d}t = h(0) + \int_0^t f(h(t), t; \theta) \mathrm{d}t, \tag{2}$$

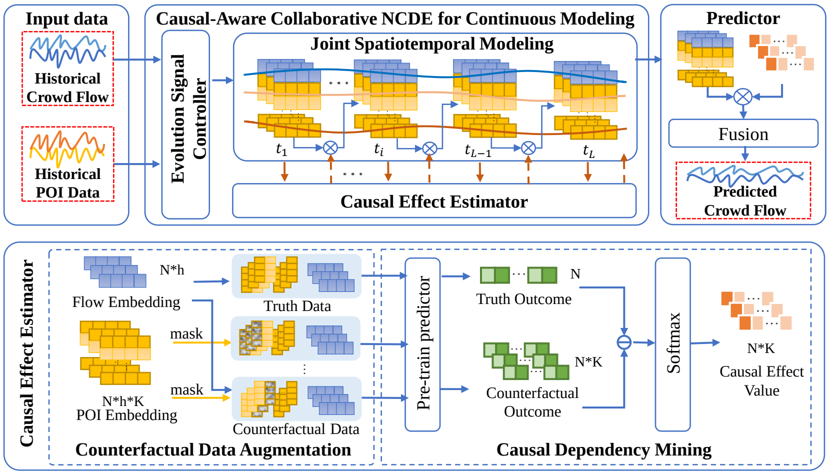


Fig. 1: Framework overview of C^{3}DE.

where a neural network $f(\cdot)$ with parameter θ parameterize the derivative of the hidden state, i.e., $\frac{\mathrm{d}h}{\mathrm{d}t} := f(h(t), t; \theta)$. The evolution process is computed using ODE solvers, such as the Euler method and Runge-Kutta. To improve efficiency, the adjoint sensitivity method is often employed to compute the parameter gradients via the adjoint equations, rather than direct backpropagation.

Neural CDEs. Neural CDEs [17] are the extension of neural ODEs. Neural CDEs introduces an external control signal X_t, which drives the evolution of the hidden state $h(t)$, making it dependent on both its own dynamics and the control signal. Specifically, it can be expressed as:

$$h(t) = h(0) + \int_0^t f(h(t), t; \theta)\mathrm{d}X_t, \tag{3}$$

where X_t is a continuous path defined in a Banach space, representing the external control signal. Different from Eq. 2, it represents a Riemann−Stieltjes integral, allowing to model the influence of control signal on system's evolution.

4 Methodology

In this section, we introduce the proposed C^{3}DE framework, as shown in Fig. 1. It comprises two main modules. The first is the main pipeline of causal-aware collaborative neural CDE, which models the continuous evolution of collaborative signals while uncovering their potential causal impacts. The second is the well-designed causal effect estimator, consisting of counterfactual data augmentation and causal dependency mining, designed to explore the causal relationships between collaborative signals.

4.1 Dual Neural CDE

We first introduce a naive dual neural CDE as $F(\cdot)$ for modeling both the crowd flow and the POI distribution data simultaneously. It is formulated as:

$$\begin{cases} h_x(t') = h_x(0) + \int_0^{t'} f(h_x(t), t; \theta)\mathrm{d}X_t, & t' \in [0, T], \\ h_p(t'') = h_p(0) + \int_0^{t''} f(h_p(t), t; \theta)\mathrm{d}P_t, & t'' \in [0, M], \end{cases} \tag{4}$$

where $h_x(t')$ and $h_p(t'')$ represent the hidden states of crowd flow and POI distribution at t' and t'' respectively, which can be computed by an adaptive step-size solver or a fixed step-size solver like Runge-Kutta and Euler methods [14]. The control signals X_t and P_t guide the dynamic evolution process of the dual neural CDE, which are derived from the urban crowd flow and the POI distribution, respectively. Given the crowd flow $X_{t'-T+1:t'} \in \mathbb{R}^{T\times N\times C}$ and POI distribution $P_{t''-M+1:t''} \in \mathbb{R}^{M\times N\times K}$, we use the natural cubic spline [5] to create continuous paths which are needed in a neural CDE for control signals:

$$X_t = Spline(X_{t'-T+1:t'}), \quad P_t = Spline(P_{t''-M+1:t''}), \tag{5}$$

where $Spline(\cdot)$ denotes the natural cubic spline function, which generates continuous, smooth, and twice-differentiable paths for a given input, ensuring accurate and stable gradient computation.

Exemplification. In Eq.(4), the function $f(\cdot)$, which deals with the spatio-temporal features of signals, can be applied with any model for processing sequential data.

Without loss of generality, we leverage Gated Recurrent Unit (GRU) [6] as an example to illustrate the derivation of $f(\cdot)$ in Eq.(4). To extend the state update of GRU to the continuous time domain, we introduce the state change $\triangle h_t$ over the time interval $\triangle t$, defined as:

$$\triangle h_t = h_t - h_{t-\triangle t} = (1 - z_t) \odot (\tilde{h}_t - h_{t-\triangle t}), \tag{6}$$

where z_t and $\tilde{h}_t$ are the intermediate vectors of GRU. As the time interval $\triangle t$ tends to 0, it can be transformed into the differential form of continuous time:

$$\frac{\mathrm{d}h(t)}{\mathrm{d}t} = (1 - z_t) \odot (\tilde{h}_t - h_{t-\triangle t}). \tag{7}$$

Similarly to the GRU model, the state update of any function $f(\cdot)$ for processing sequential data can be extended to the continuous time domain.

4.2 Causal-aware Collaborative Neural CDE (C^{3}DE)

Intuitively, there is a tight interaction between long-term evolution of the urban crowd flow and that of the POI distribution. These mutual causalities result in the insufficient modeling in the dual neural CDE which deals with the two collaborative signals respectively. To explore causal impacts of collaborative signals, we integrate the causal awareness mechanism and propose the causal-aware collaborative neural CDE. Regarding the accumulation of spurious correlations during continuous evolution, we employ a dynamic correction mechanism in $F(\cdot)$ and rewrite the Eq.(4) to alleviate spurious correlations as follows:

$$\begin{cases} h_x(t') = h_x(0) + \int_0^{t'} f(h_x(t), t; \theta)\mathrm{d}X_t, & t' \in [0, T], \\ h_p(t'') = h_p(0) + \int_0^{t''} \mathcal{C} \cdot f(h_p(t), t; \theta)\mathrm{d}P_t, & \mathcal{C} \in [\mathcal{C}_1, ..., \mathcal{C}_L], t'' \in [0, M]. \end{cases} \tag{8}$$

where $\mathcal{C} \in \mathbb{R}^{N \times K}$ is the causal impact weight, used to correct the biased POI representations, and L is the number of observation points during the evolution. We design a counterfactual-based causal effect estimator $g(\cdot)$ to compute $\mathcal{C}$. Specifically, it takes the hidden states of collaborative signals at observation points during the evolution as input to quantify:

$$\mathcal{C}_i = g(h_x(t'_i), h_p(t''_i)), i \in \{1, ..., L\}, \tag{9}$$

where $\{t'_1, ..., t'_i, ..., t'_L\} \subset [0, T]$ and $\{t''_1, ..., t''_i, ..., t''_L\} \subset [0, M]$ are the evenly spaced observation time points within their intervals. $\mathcal{C}_i$ denotes the causal impact weight of i-th observation point. Next, we introduce the design of $g(\cdot)$.

Counterfactual-Based Causal Effect Estimator. Inspired by the success of counterfactual data augmentation in natural language processing [35] and dynamic system [28], we explore the causal impact of POI on crowd flow from a counterfactual perspective, which aims to answer: "How would crowd flow change if a certain POI were changed?".

Specifically, the causal effect estimator consists of two modules: counterfactual data augmentation and causal dependency mining.

Counterfactual Data Augmentation. To answer the above question, we propose a counterfactual data augmentation method based on category-level perturbation, simulating various scenarios of POI changes. Specifically, given the POI representation $h_p(t''_i) \in \mathbb{R}^{N \times K \times H}$ at the i-th observation points, where H denotes the hidden space dimension, the counterfactual data for the k-th POI category is constructed as $h^k_{p^*}(t''_i)$:

$$h^k_{p^*}(t''_i) = h_p(t''_i) \odot M_k, \tag{10}$$

where M_k is a perturbation matrix, such as zero-setting, random perturbation, or mean replacement, that controls the category-level perturbation on the k-th POI category. Take the zero-setting perturbation as an example, the perturbation matrix M_k can be defined as:

$$(M_k)_{n,j,h} = \begin{cases} 0, & \text{if } j = k \\ 1, & \text{otherwise,} \end{cases} \tag{11}$$

where n is the region index, j is the POI category index, and h is the hidden space dimension index.

We can generate a set of counterfactual POI data by applying perturbations to the K POI categories: $\left\{h_{p^*}^1(t_i''), ..., h_{p^*}^k(t_i''), ..., h_{p^*}^K(t_i'')\right\}$, where each represents the POI representation under a specific POI category's perturbation. Next, we pair the generated counterfactual POI representation $h_{p^*}^k(t_i'')$ with the crowd flow representation $h_x(t_i')$ to obtain K pairs of counterfactual samples $\mathfrak{D}_{cf}$:

$$\mathfrak{D}_{cf} = \left\{\left(h_x(t_i'), h_{p^*}^k(t_i'')\right) \mid h_{p^*}^k(t_i'') \in \left\{h_{p^*}^1(t_i''), ..., h_{p^*}^K(t_i'')\right\}\right\}, \tag{12}$$

the unperturbed POI representation $h_p(t_i'')$ and $h_x(t_i')$ are paired to form the factual sample $\mathfrak{D}_{fact}$:

$$\mathfrak{D}_{fact} = \{(h_x(t_i'), h_p(t_i''))\}. \tag{13}$$

Causal Dependency Mining. To evaluate the dynamic impacts of collaborative signals and reveal their causal dependency, we propose a causal dependency mining module based on factual and counterfactual samples.

We first pre-train a predictor $\mathcal{T}(\cdot)$ with the loss function $\ell(\cdot)$. Notably, $\mathcal{T}(\cdot)$ can be any spatiotemporal model, and here we use MTGNN [29] as the backbone:

$$\mathcal{T}^* = arg\min_{\mathcal{T}} \ell(\mathcal{T}(X_{t-T+1:t}, P_{t'-M+1:t'}), X_{t+1:t+S}). \tag{14}$$

Next, we sequentially input the counterfactual samples into the predictor $\mathcal{T}(\cdot)$ to obtain the counterfactual output O_{x,p^*}^k:

$$O_{x,p^*}^k = \mathcal{T}(h_x(t_i'), h_{p^*}^k(t_i'')) \in \mathbb{R}^N, \quad \forall (h_x(t_i'), h_{p^*}^k(t_i'')) \in \mathfrak{D}_{cf}. \tag{15}$$

Meanwhile, we input the factual samples $\{(h_x(t_i'), h_p(t_i''))\} \in \mathfrak{D}_{fact}$ into the same $\mathcal{T}(\cdot)$ to obtain the factual output $O_{x,p}^k$, which is considered as an anchor:

$$O_{x,p}^k = \mathcal{T}(h_x(t_i'), h_p(t_i'')) \in \mathbb{R}^N. \tag{16}$$

We quantify the causal impact of a specific POI category on crowd flow by the absolute difference between the anchor and counterfactual outputs. For the k-th category of POI, the causal effect value is computed as followed:

$$\mathcal{C}_k(i) = |O_{x,p}^k - O_{x,p^*}^k| \in \mathbb{R}^N. \tag{17}$$

A larger causal effect value $\mathcal{C}_k(i)$ indicates significant fluctuations in crowd flow with the changes in k-th POI category, suggesting its key role in flow variation.

To evaluate the causal impacts of all categories, we apply $Softmax$ function to normalize all causal effect values, obtaining the overall causal effect value at i-th observation point:

$$\mathcal{C}(i) = Softmax(\mathcal{C}_1(i), ..., \mathcal{C}_K(i)) \in \mathbb{R}^{N \times K}. \tag{18}$$

By performing the above operation at L observation points, we can capture the dynamic causal impacts of POI on crowd flow.

4.3 Predictor and Overall Objective

Predictor. Through the modeling process of C^{3}DE, we obtain the crowd flow representation $h_x(t')$ and POI distribution representation $h_p(t'')$, capturing historical evolution and the key causal features for the prediction task. We fuse their representations to explore the evolution of collaborative signals between POI and crowd flow, resulting in a comprehensive representation H that captures the multidimensional features of crowd flow changes:

$$H = \sigma(h_x(t') \cdot W_x + b_x) \odot (h_p(t'') \cdot W_p + b_p), \tag{19}$$

where W_x and W_p are the learnable weight matrices, b_x and b_p are learnable bias, and $\sigma(\cdot)$ denotes the *sigmod* function.

Subsequently, the fused representation H is fed into a multilayer perceptron-based predictor $G(\cdot)$ to predict the next S time steps, as shown below:

$$\hat{X}_{t+1:t+S} = G(H; \theta_g) \in \mathbb{R}^{S \times N \times C}, \tag{20}$$

where $\hat{X}_{t+1:t+S}$ denotes the predicted values.

Overall Objective. Finally, we adopt the Huber loss as the objective function $\ell(\cdot)$. Compared to the traditional squared error loss, it exhibits greater robustness in handling outliers. For simplicity, we use Y and $\hat{Y}$ to represent $X_{t'+1:t'+S}$ and $\hat{X}_{t'+1:t'+S}$, respectively. The learning objective is expressed as:

$$\ell(Y, \hat{Y}) = \begin{cases} \frac{1}{2}(Y - \hat{Y}), & |Y - \hat{Y}| \leq \delta \\ \delta|Y - \hat{Y}| - \frac{1}{2}\delta^2, & \text{otherwise}, \end{cases} \tag{21}$$

where δ is a hyperparameter that controls the sensitivity to outliers.

4.4 Complexity Analysis of C^{3}DE

In the solving process of C^{3}DE, we adopt the adjoint sensitivity method [8] to compute gradients efficiently. Unlike traditional backpropagation, this method solves an auxiliary adjoint differential equation to trace gradients backward in time, requiring only the storage of the final state rather than the entire forward trajectory. This leads to a space complexity of $O(N \cdot d)$, where N denotes the number of nodes and d is the dimension of the hidden state, significantly lower than that of standard backpropagation. However, this advantage in space comes at the cost of additional computation time. Since the adjoint method requires an extra backward integration, the time complexity is approximately $O(2 \cdot N_{fe} \cdot C_f)$, where N_{fe} is the number of times the CDE solver calls the function $f(\cdot)$, and C_f is the time cost of the spatio-temporal modeling function $f(\cdot)$. Given that our task focuses on long-term crowd flow prediction, where prediction accuracy and stability are prioritized over real-time inference, this trade-off in computation cost is acceptable. The advantages in storage space and model performance make our approach both practical and deployable in real-world urban management applications.

Table 1: Statistics of urban crowd flow dataset.

Description	NYC-1	NYC-2	Beijing
time spanning	2012.06∼ 2014.05	2014.09∼2016.12	2018.07 ∼ 2019.10
# of time steps	16,128	17,424	10,241
# of records	2,322,432	2,787,840	1,894,585
# of nodes	144	160	185

Table 2: Statistics of POI distribution dataset.

Description	NYC-1	NYC-2	Beijing
time spanning	2011.10∼ 2014.05	2014.01∼2016.12	2017.10 ∼ 2019.10
# of records	23,040	28,800	32,375
# of nodes	144	160	185
# of types	5	5	7

5 Experiments

5.1 Experimental setup

Dataset. We evaluate the proposed framework on three real-world urban crowd flow datasets and their corresponding POI datasets: *NYC-1* and *NYC-2*, collected from NYC OpenData[1], and *Beijing* [34]. We summarize the statistics for three datasets in Table 1 and Table 2.

Baselines. To evaluate the effectiveness of our C^{3}DE, we compare it with the following baselines:

- Traditional methods: **HA** predicts future values by averaging historical data from the same time period. **SVR** is a regression method based on support vector machines.
- Discrete methods: **STGCN** [31] learns spatio-temporal dependencies with a graph convolutional structure. **GWNET** [30] uses an adaptive adjacency matrix to capture hidden spatial dependencies. **STSGCN** [26] captures localized correlations via the synchronous mechanism. **MTGNN** [29] is a general GNN for modeling multivariate time series. **STWave** [9] is a decomposition-based framework that decouples flow using wavelet transform.
- Continuous methods: **STGODE** [10] extends GNNs with tensor-based ODEs to build deeper networks. **STG-NCDE** [5] designs two NCDEs to model temporal and spatial dependencies. **MTGODE** [16] uses NODEs and dynamic graph structure learning to model continuous dynamics.

Evaluation Metrics. We use Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) to evaluate performance. Lower values of these metrics indicate better performance.

Implementation Details. We implemented C^{3}DE in PyTorch using the Adam optimizer with a learning rate $lr = 0.001$, weight decay of 5×10^{-4}, and batch

[1] https://opendata.cityofnewyork.us/

Table 3: Overall performance comparison on three real-world datasets. Highlighting denotes the best results and **bolding** denotes the second-best results.

Dataset	Method	Horizon 7			Horizon 14			Average		
		MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
Beijing	HA	289.40	756.21	81.4%	289.40	756.21	81.4%	289.40	756.21	81.4%
	VAR	280.41	724.90	74.2%	285.71	731.12	79.5%	283.06	728.01	76.8%
	STGCN	271.55	669.84	77.3%	283.85	692.63	80.1%	277.70	681.24	78.7%
	GWNET	170.33	414.13	**39.0**%	208.77	503.38	47.7%	189.55	458.76	43.4%
	STSGCN	198.82	507.28	50.7%	221.04	546.77	53.3%	209.93	527.02	52.0%
	MTGNN	196.19	483.25	38.9%	232.93	561.09	48.2%	214.56	522.17	43.6%
	STWave	212.40	522.14	48.6%	242.89	588.74	54.4%	227.65	555.44	51.5%
	STGODE	209.26	498.13	58.1%	234.87	560.02	59.2%	222.07	529.08	58.7%
	STG-NCDE	**159.56**	**404.01**	39.2%	**202.23**	**491.81**	**47.0**%	**180.89**	**447.91**	**43.1**%
	MTGODE	218.85	572.61	75.1%	222.38	594.06	76.4%	220.61	583.33	75.7%
	C^{3}DE	117.56	245.20	35.3 %	124.05	248.39	41.9 %	120.81	246.80	38.6 %
NYC-1	HA	39.065	106.79	34.42%	39.065	106.79	34.42%	39.065	106.79	34.42%
	VAR	36.186	101.18	32.70%	37.393	104.03	33.80%	36.789	102.60	33.25%
	STGCN	28.777	63.783	26.68%	30.316	77.264	26.99%	29.546	70.524	26.84%
	GWNET	25.060	60.958	17.40%	24.970	61.039	17.96%	25.015	60.999	17.68%
	STSGCN	26.143	62.917	26.19%	26.943	63.592	26.14%	26.543	63.255	26.17%
	MTGNN	24.165	57.958	18.94%	24.899	58.121	20.30%	24.532	58.040	19.62%
	STWave	24.354	57.839	16.90%	24.824	58.256	17.35%	24.589	58.047	17.13%
	STGODE	24.868	58.664	20.01%	25.095	58.777	20.75%	24.982	58.721	20.38%
	STG-NCDE	24.693	58.289	19.53%	24.988	58.479	20.52%	24.841	58.384	20.03%
	MTGODE	**24.141**	**57.771**	**16.37**%	**24.758**	**57.726**	**17.06**%	**24.449**	**57.748**	**16.72**%
	C^{3}DE	23.997	55.598	14.16 %	24.050	56.208	14.31 %	24.024	55.903	14.24 %
NYC-2	HA	24.963	63.686	24.73%	24.963	63.686	24.73%	24.963	63.686	24.73%
	VAR	23.908	61.967	16.46%	24.232	62.761	16.22%	24.070	62.364	16.34%
	STGCN	19.969	34.225	11.63%	20.619	36.281	11.64%	20.294	35.253	11.64%
	GWNET	11.043	27.509	8.73%	11.665	29.120	8.90%	11.354	28.314	8.82%
	STSGCN	11.424	27.697	8.84%	11.736	29.163	8.92%	11.580	28.430	8.88%
	MTGNN	10.699	26.470	**8.18**%	11.102	28.049	8.36%	10.901	27.260	8.27%
	STWave	10.933	26.074	8.37%	11.676	28.056	8.45%	11.304	27.065	8.41%
	STGODE	12.780	28.357	11.55%	12.780	29.633	11.43%	12.780	28.995	11.49%
	STG-NCDE	10.876	27.601	8.22%	11.342	28.970	8.78%	11.109	28.286	8.50%
	MTGODE	**10.545**	**26.028**	**8.18**%	**10.816**	**28.023**	**8.33**%	x**10.681**	**27.026**	**8.26**%
	C^{3}DE	10.159	25.619	7.98 %	10.115	27.257	8.07 %	10.137	26.438	8.03 %

size $B = 64$. The representation size was fixed to 64 for all methods. We set the historical observation length to $T = 14$, $M = 4$, and the future prediction length to $S = 14$. For the Beijing, NYC-1 and NYC-2 datasets, we set the number of observation points to $L = 10/8/8$, respectively. For counterfactual data augmentation, we applied zero-setting perturbation by default. We used an adaptive solver for the Beijing and NYC-2 datasets and the 4th order Runge-Kutta (RK4) solver with a step size of 1.2 for NYC-1. The codes are available at https://github.com/Sonder-arch/C3DE.

5.2 Overall Performance

We evaluate C^{3}DE on three real-world datasets for the task of long-term urban crowd flow prediction, with results in Table 3. We observe: (1) Statistical

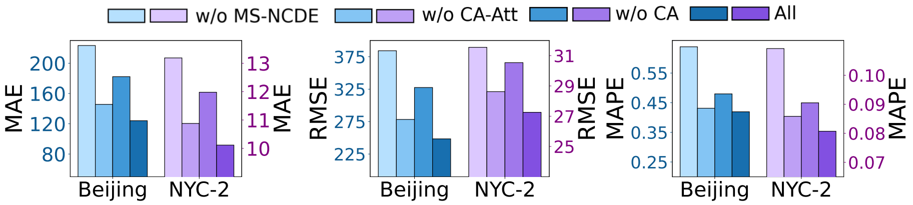


Fig. 2: Ablation study on Beijing and NYC-2 datasets.

methods HA and VAR perform the worst, as relying solely on historical data fails to capture complex and dynamic spatiotemporal patterns, leading to significant prediction errors. (2) MTGODE performs sub-optimally on the NYC-1 and NYC-2 datasets but experiences a sharp performance drop on the Beijing dataset. While its continuous-time modeling and dynamic graph structure can effectively capture long-term dependencies in the stable NYC data, it struggles with the complex temporal dynamics of the more volatile Beijing dataset, resulting in instability. In contrast, STG-NCDE achieves the second-best performance on the Beijing dataset, likely due to its NCDEs-based independent spatio-temporal modeling, which better captures sudden flow changes and intricate temporal dynamics. (3) Continuous methods do not always outperform discrete methods. MTGNN consistently surpasses STGODE across all three datasets, likely because while STGODE employs a continuous GNN with residual connections to avoid over-smoothing, it still relies on a fixed graph structure, limiting its ability to capture potential correlations. MTGNN overcomes this limitation with node-adaptive graph convolution. (4) C^{3}DE consistently outperforms all baselines, especially on the Beijing dataset, demonstrating its superior generalization and stability. It is due to its ability to uncover complex data changes through counterfactual inference. When handling highly volatile collaborative signals, it more accurately models their continuous evolution, demonstrating stronger robustness and generalization in complex scenarios.

5.3 Ablation Study

In this section, we further validate the effectiveness of the proposed modules in C^{3}DE, with a particular focus on the continuous modeling and causal mining modules. Specifically, we design the following variants, and the experimental results on Beijing and NYC-2 datasets are shown in Fig. 2.

- *w/o MS-NCDE*: Remove the NCDE continuous modeling module.
- *w/o CA-Att*: Replacing counterfactual-based causal effect values with attention mechanism-based values.
- *w/o CA*: Remove the counterfactual-based causal effect estimator module totally, only simply fuse POI and flow final representations.
- *All*: It is our complete framework.

Effectiveness of Dynamic Continuous Modeling. Experimental results show that *w/o MS-NCDE* performs the worst across the three datasets, high-

Table 4: The impact of different counterfactual strategies on Beijing dataset.

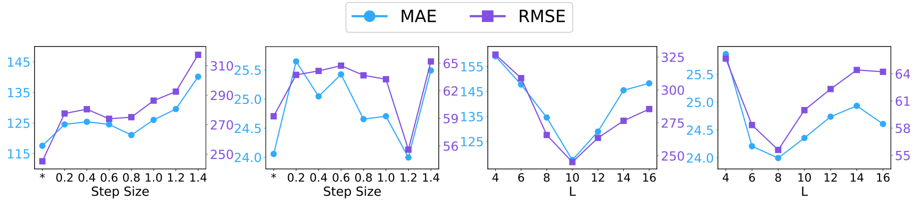
Method	**Horizon 7**			**Horizon 14**			**Average**		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
baseline (*w/o CA*)	160.66	291.87	41.8%	182.26	327.58	47.9%	171.46	309.73	44.9%
C^{3}DE-*random*	**123.12**	**252.36**	**35.9%**	**136.08**	**271.18**	**42.7%**	**129.60**	**261.77**	**39.3%**
C^{3}DE-*zero*	117.56	245.20	35.3%	124.04	248.39	41.9%	120.81	246.80	38.6%
C^{3}DE-*mean*	128.53	266.86	36.2%	142.66	278.08	43.4%	135.60	272.47	39.8%

lighting the effectiveness of collaborative NCDE in continuous collaborative signals modeling. Specifically, collaborative NCDE, formulated as differential equations, can smoothly capture the fine-grained continuous spatio-temporal evolution of collaborative signals, thereby effectively learning potential changes beyond the observation points.

Effectiveness of Causal Dependency Mining. From the results, we can see that the *All* outperforms the *w/o CA* and *w/o CA-Att*, which demonstrates the effectiveness of the counterfactual-based causal effect estimator in capturing and eliminating spurious correlations. Further, we also find that: first, the *w/o CA* variant performs worst among the three variants, indicating that relying solely on POI distribution for prediction is insufficient. While POI distribution can partially reflect flow dynamics, not all POI have a substantive causal relationship with crowd flow. Many POIs exhibit only superficial correlations, which introduce spurious relationships and weaken the model's expressiveness, leading to the performance degradation of *w/o CA*. Second, *w/o CA-Att* models collaborative signals based on attention, dynamically assigning weights to POI distributions to highlight key signals. However, it fundamentally relies on data correlations, making it challenging to distinguish true causal relationships. Third, *All* employs a counterfactual framework for causal inference, capturing more interpretable causal dependencies and mitigating spurious correlations, leading to superior modeling of collaborative signal evolution.

5.4 The Impact of Different Counterfactual Strategies

In this section, we explore the impact of different counterfactual strategies on prediction performance. Specifically, we employ three strategies: "*random*", "*zero*", and "*mean*", against the baseline *w/o CA*, which removes the causal effect estimator module. Table 4 presents the results on Beijing dataset, leading to the following findings: (1) The "*zero*" strategy performs best. As a stringent intervention, it sets the target POI representation to zero, effectively removing its feature information to explore its direct causal impact on crowd flow. (2) Unlike "*zero*", the "*random*" strategy introduces random noise to replace the target POI representation. However, this may introduce uncertainty into the model's causal inference process, leading to suboptimal performance. (3) The "*mean*" strategy averages all POIs representations except the target and uses this average as its counterfactual representation. However, it achieves the lowest performance, possibly because the averaged spatial distribution information blurs the target



(a) Beijing. (b) NYC-1. (c) Beijing. (d) NYC-1.

Fig. 3: The impacts of *Step Size* and L on two datasets.

POI's unique causal effect, making it challenging for the model to capture its true impact. (4) Notably, all three strategies outperform the baseline "*w/o CA*", demonstrating that our framework effectively mines the true causality and thus enhances performance regardless of the intervention strategy.

5.5 Impacts of Hyper-Parameters

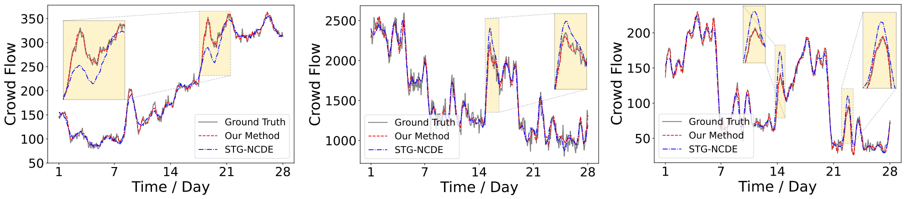
We conduct experiments to validate the impacts of different solvers and the number of observation points L in the dynamic correction mechanism. First, we choose the adaptive solver and the commonly used fixed-step RK4 solver with different step sizes. Note that, smaller step sizes yield finer data fitting. As illustrated in Fig. 3a and Fig. 3b, the adaptive solver (marked with *) and RK4 solver with a step size of 1.2 perform best on the Beijing and NYC-1 datasets, respectively. Performance improves as the step size decreases, as larger step sizes hinder the model to capture precise dynamic changes. However, beyond a threshold, further reducing the step size offers no gains, as the variation between each step becomes negligible. Second, the impacts of varying L from 4 to 16 are shown in Fig. 3c and Fig. 3d. The best results are achieved with $L = 10$ for Beijing and $L = 8$ for NYC-1. A small L limits the model's ability to detect and reduce spurious correlations between collaborative signals, while a large L hinders its capacity to capture dynamic signal variations.

5.6 Visualization of Prediction Results

We conduct a case to demonstrate the advantages of our method over the continuous modeling baseline, STG-NCDE. Specifically, C^{3}DE excels at capturing the early-stage changes in fluctuations, which are critical for accurate prediction. As shown in Fig. 4, it can not only accurately identify the growth trend at the beginning of the fluctuation(Fig. 4a) but also capture the subsequent decline(Fig. 4b and Fig. 4c), which is due to C^{3}DE's ability to accurately model the relationships between collaborative signals. These results indicate that C^{3}DE effectively captures the dynamic changes in the system, precisely tracks the early stages of fluctuations, and accurately predicts the future flow variation trends.

6 Conclusion

In this paper, we proposed C^{3}DE, a framework with causal-aware collaborative neural controlled differential equations for long-term urban crowd flow prediction. We first introduced the neural CDE with a dual-path architecture to



(a) Node 50 in Beijing. (b) Node 76 in Beijing. (c) Node 130 in Beijing.

Fig. 4: Visualization of prediction results on the Beijing dataset.

capture the asynchronous dynamic evolution of collaborative signals. Next, we designed a counterfactual inference-based causal effect estimator to simulate urban dynamics under different POI distribution scenarios and mine the direct impact of different POIs on crowd flow. Moreover, we incorporated causal effect values into neural CDE. By introducing a causal effect-based dynamic correction mechanism, C^{3}DE can mitigate the accumulation of spurious correlations among collaborative signals. Extensive experiments on three real-world datasets demonstrated the significant superiority of C^{3}DE. Future work will focus on enhancing causal relationship mining efficiency by integrating causal priors based on domain knowledge in urban dynamics.

Acknowledgments. This work was supported by the Natural Science Foundation of China (Grant No.62072235) and the Young Scientists Fund of the Natural Science Foundation of Jiangsu Province, China (Grant No.BK20241402). We would also like to thank the anonymous reviewers for giving us useful and constructive comments. Additionally, we are grateful to the community and everyone who made their datasets and source codes publicly accessible. These datasets and source codes are valuable and have greatly facilitated this research.

References

1. Bai, L., Yao, L., Li, C., Wang, X., Wang, C.: Adaptive graph convolutional recurrent network for traffic forecasting. Advances in neural information processing systems **33**, 17804–17815 (2020)
2. Chen, C., Liu, Y., Chen, L., Zhang, C.: Multivariate traffic demand prediction via 2d spectral learning and global spatial optimization. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 72–88. Springer (2024)
3. Chen, G., Li, J., Lu, J., Zhou, J.: Human trajectory prediction via counterfactual analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9824–9833 (2021)
4. Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. Advances in neural information processing systems **31** (2018)
5. Choi, J., Choi, H., Hwang, J., Park, N.: Graph neural controlled differential equations for traffic forecasting. In: Proceedings of the AAAI conference on artificial intelligence. pp. 6367–6374 (2022)

6. Dey, R., Salem, F.M.: Gate-variants of gated recurrent unit (gru) neural networks. In: 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS). pp. 1597–1600. IEEE (2017)
7. Diks, C., Panchenko, V.: A new statistic and practical guidelines for nonparametric granger causality testing. Journal of Economic Dynamics and Control **30**(9-10), 1647–1669 (2006)
8. Errico, R.M.: What is an adjoint model? Bulletin of the American Meteorological Society **78**(11), 2577–2592 (1997)
9. Fang, Y., Qin, Y., Luo, H., Zhao, F., Xu, B., Zeng, L., Wang, C.: When spatio-temporal meet wavelets: Disentangled traffic forecasting via efficient spectral graph attention networks. In: 2023 IEEE 39th International Conference on Data Engineering (ICDE). pp. 517–529. IEEE (2023)
10. Fang, Z., Long, Q., Song, G., Xie, K.: Spatial-temporal graph ode networks for traffic flow forecasting. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining. pp. 364–373 (2021)
11. Gravina, A., Zambon, D., Bacciu, D., Alippi, C.: Temporal graph odes for irregularly-sampled time series. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (2024)
12. Huang, Z., Sun, Y., Wang, W.: Learning continuous system dynamics from irregularly-sampled partial observations. Advances in Neural Information Processing Systems **33**, 16177–16187 (2020)
13. Huang, Z., Sun, Y., Wang, W.: Coupled graph ode for learning interacting system dynamics. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining. pp. 705–715 (2021)
14. Islam, M.A., et al.: A comparative study on numerical solutions of initial value problems (ivp) for ordinary differential equations (ode) with euler and runge kutta methods. American Journal of computational mathematics **5**(03), 393 (2015)
15. Jiang, J., Han, C., Zhao, W.X., Wang, J.: Pdformer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction. In: Proceedings of the AAAI conference on artificial intelligence. pp. 4365–4373 (2023)
16. Jin, M., Zheng, Y., Li, Y.F., Chen, S., Yang, B., Pan, S.: Multivariate time series forecasting with dynamic graph neural odes. IEEE Transactions on Knowledge and Data Engineering **35**(9), 9168–9180 (2022)
17. Kidger, P., Morrill, J., Foster, J., Lyons, T.: Neural controlled differential equations for irregular time series. Advances in neural information processing systems **33**, 6696–6707 (2020)
18. Li, H., Gu, J., Lu, X., Shen, D., Liu, Y., Deng, Y., Shi, G., Xiong, H.: Beyond relevance: Factor-level causal explanation for user travel decisions with counterfactual data augmentation. ACM Transactions on Information Systems **42**(5), 1–31 (2024)
19. Li, H., Gu, J., Ying, H., Lu, X., Yang, J.: User multi-behavior enhanced poi recommendation with efficient and informative negative sampling. In: Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data. pp. 149–165. Springer (2022)
20. Liang, Y., Ke, S., Zhang, J., Yi, X., Zheng, Y.: Geoman: Multi-level attention networks for geo-sensory time series prediction. In: IJCAI. vol. 2018, pp. 3428–3434 (2018)
21. Liu, Z., Ding, J., Zheng, G.: Frequency enhanced pre-training for cross-city few-shot traffic forecasting. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 35–52. Springer (2024)

22. Long, Q., Fang, Z., Fang, C., Chen, C., Wang, P., Zhou, Y.: Unveiling delay effects in traffic forecasting: A perspective from spatial-temporal delay differential equations. In: Proceedings of the ACM on Web Conference 2024. pp. 1035–1044 (2024)
23. Rong, C., Li, T., Feng, J., Li, Y.: Inferring origin-destination flows from population distribution. IEEE Transactions on Knowledge and Data Engineering **35**(1), 603–613 (2021)
24. Shao, X., Wang, H., Chen, X., Zhu, X., Zhang, Y.: Cube: Causal intervention-based counterfactual explanation for prediction models. IEEE Transactions on Knowledge and Data Engineering (2023)
25. Shao, Z., Zhang, Z., Wei, W., Wang, F., Xu, Y., Cao, X., Jensen, C.S.: Decoupled dynamic spatial-temporal graph neural network for traffic forecasting. Proceedings of the VLDB Endowment **15**(11), 2733–2746 (2022)
26. Song, C., Lin, Y., Guo, S., Wan, H.: Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In: Proceedings of the AAAI conference on artificial intelligence. pp. 914–921 (2020)
27. Tian, B., Cao, Y., Zhang, Y., Xing, C.: Debiasing nlu models via causal intervention and counterfactual reasoning. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 11376–11384 (2022)
28. Wang, Z., Zhang, J., Xu, H., Chen, X., Zhang, Y., Zhao, W.X., Wen, J.R.: Counterfactual data-augmented sequential recommendation. In: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. pp. 347–356 (2021)
29. Wu, Z., Pan, S., Long, G., Jiang, J., Chang, X., Zhang, C.: 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. pp. 753–763 (2020)
30. Wu, Z., Pan, S., Long, G., Jiang, J., Zhang, C.: Graph wavenet for deep spatial-temporal graph modeling. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence. pp. 1907–1913 (2019)
31. Yu, B., Yin, H., Zhu, Z.: Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. pp. 3634–3640 (2018)
32. Yu, C., Wang, F., Shao, Z., Sun, T., Wu, L., Xu, Y.: Dsformer: A double sampling transformer for multivariate time series long-term prediction. In: Proceedings of the 32nd ACM international conference on information and knowledge management. pp. 3062–3072 (2023)
33. Zeng, J., Zhang, G., Rong, C., Ding, J., Yuan, J., Li, Y.: Causal learning empowered od prediction for urban planning. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. pp. 2455–2464 (2022)
34. Zheng, Z., Gu, J., Zhou, Q., Lu, X.: Prediction in long-term evolution: Exploiting the interaction between urban crowd flow variation and poi transition patterns. In: 2023 IEEE International Conference on Data Mining (ICDM). pp. 1559–1564. IEEE (2023)
35. Zmigrod, R., Mielke, S.J., Wallach, H., Cotterell, R.: Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 1651–1661 (2019)