

CASTANET: Causality-Aware Spatio-Temporal Adversarial Network Using Traffic Incident Effects

Toshiya Kitahara, Ryu Shirakami, Koh Takeuchi, Hisashi Kashima

Abstract—Predicting non-periodic traffic congestion caused by sudden incidents (e.g., accidents and road damage) is crucial for advanced intelligent transportation systems. However, incident-driven congestion is difficult to forecast because incidents are extremely sparse, occur at specific times and locations, and have heterogeneous impacts depending on the traffic context. While recent deep learning approaches have significantly improved periodic traffic forecasting, their performance on non-periodic congestion remains limited, partly because incident records are not explicitly incorporated and their occurrence is strongly biased in space and time. To address these challenges, we propose CASTANET, which integrates spatio-temporal graph neural networks and causal treatment effect estimation to utilize incident records while mitigating selection bias. Experiments on real-world traffic data and accident records from Tokyo, which we treat as incidents, show that CASTANET reduces RMSE by 4.0% overall compared to the best baseline and by 10.1% on incident-conditioned evaluation, with gains reaching 14.55% under severe congestion.

I. INTRODUCTION

Traffic congestion management is an essential function of advanced intelligent transportation systems (ITS). In urban road networks, congestion can be broadly categorized into periodic and non-periodic congestion [1]. Periodic congestion arises from regular social activities such as commuting and logistics, and therefore repeatedly occurs at similar times and locations. In contrast, non-periodic congestion occurs irregularly and may propagate rapidly through the road network because sudden incidents, such as traffic accidents or road damage, often trigger such congestion. If the growth of such incident-driven congestion can be forecast, traffic operators can respond proactively, for example by adapting signal timings and allocating operational resources. We also need to know whether the growth is mainly attributable to the incident or would have occurred even without it.

Although the causes of non-periodic congestion are varied, we hypothesize that incident-driven congestion growth is predictable by modeling the interaction between traffic conditions, road structure, and incidents. Recently, spatio-temporal graph neural networks (STGNNs) [2], [3], [4] have shown their potential to capture such complex relationships from large-scale traffic observations. However, predicting non-periodic congestion remains challenging for two key reasons.

First, the impact of a sudden incident is heterogeneous [5]. We show traffic observations around incidents on two different roads in Figure 1. In the left figure, an incident at around

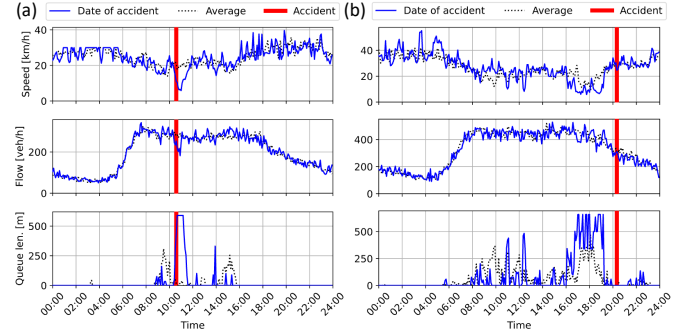


Fig. 1. Speed, flow, and queue length under two incidents. (a) Speed and flow drop while queue length rises after the incident. (b) No clear traffic changes are observed.

10:30 causes a sharp decline in speed and flow, leading to a rapid increase in queue length. In contrast, in the right figure, there are no apparent changes in traffic after the incident. These contrasting outcomes arise because the impact of an incident depends on the incident itself, traffic conditions, and the local road network structure. A predictive model must therefore capture these context-dependent interactions rather than treating all incidents uniformly.

Second, incident data are inherently sparse and exhibit strong spatio-temporal biases [6]. In our traffic dataset from Tokyo, incidents occurred in only 0.01% of all samples. The fact that the locations and times where incidents occur are not uniformly distributed is an additional complexity. Figure 2 shows spatial and temporal histograms of incidents in Tokyo, indicating non-uniform distribution across time or space. This sparsity makes prediction difficult, but a more subtle problem arises: incidents occur more frequently under specific conditions. Naively adding incident indicators as input features to an STGNN therefore risks confounding the causal effect of an incident with the traffic conditions that make incidents more likely. A model may attribute congestion to an incident even when similar congestion would have occurred without it, which is undesirable for ITS decision-making.

To mitigate this confounding, we draw on the treatment effect estimation framework from causal inference [7], where an incident can be viewed as a non-random intervention and the impact on congestion is the treatment effect. A key technique developed in this framework is balanced representation learning [8], which learns latent representations where treated and control groups are similarly distributed, thereby reducing selection bias and improving predictive accuracy even

Toshiya Kitahara is with Sumitomo Electric System Solutions Co., Ltd. Ryu Shirakami, Koh Takeuchi and Hisashi Kashima are with Kyoto University. Corresponding author: Toshiya Kitahara (kitahara-toshiya@seiss.co.jp).

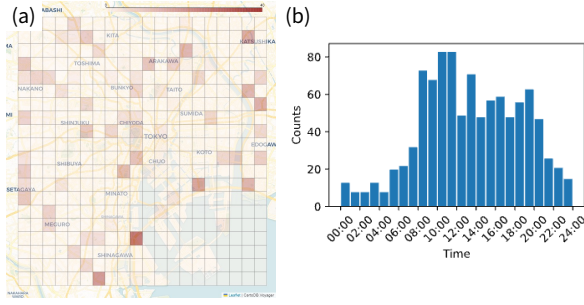


Fig. 2. (a) The spatial distribution of incidents in Tokyo from May 1 to September 26, where darker meshes indicate higher frequencies. (b) Temporal histogram of incidents.

for factual outcomes. These methods have been extended to multivariate time series [9], [10] and spatio-temporal data [11], [12]. However, these methods have been developed for domains such as epidemiology [11], and hierarchical spatial systems [12], and have not been applied to traffic congestion prediction, where incidents are extremely sparse, exhibit strong spatio-temporal selection bias, and interact with directional traffic propagation on road networks.

In this paper, we propose a causality-aware spatio-temporal adversarial network (CASTANET) for Incident-Aware traffic congestion prediction. CASTANET predicts future congestion while accounting for traffic context, including traffic history, road structure, and records of sudden traffic incidents. The key idea is to integrate STGNNs with treatment effect estimation, so that sudden incidents can be treated as non-random interventions that affect future traffic. First, we combine an STGNN with balanced representation learning methods [9] to learn latent representations of traffic conditions from historical traffic observations, incident history, and road structure [2]. Second, we predict future congestion using the learned representations with the current presence or absence of a sudden incident. To mitigate bias due to spatio-temporally imbalanced incident occurrences, we incorporate adversarial representation learning (ARL) [9], [10], which encourages balanced representations and reduces dependence on factors that merely make incidents more likely to occur.

We conduct experiments on a real-world dataset consisting of traffic measurements and incident records for approximately two thousand road segments in Tokyo, Japan. We focus on predicting queue length, which is an important quantity for traffic signal control. Experimental results show that CASTANET reduces queue length prediction RMSE by 4.0% overall compared to the best external baseline (QTNet) and by 10.1% in post-incident RMSE compared to the best ablation variant, with gains reaching 14.55% in severe congestion against state-of-the-art models.

We summarize the contributions of this study as follows:

- **Causality-aware formulation of incident-driven congestion prediction:** We formulate Incident-Aware traffic prediction within the treatment effect estimation framework, leveraging balanced representation learning to re-

duce the influence of confounding factors correlated with both incidents and congestion.

- **Integration of STGNNs with adversarial balanced representations:** We develop CASTANET, which combines an STGNN encoder with adversarial balanced representation learning tailored to the challenges of traffic incident data, including extreme sparsity and strong spatio-temporal selection bias in incident occurrences.
- **Empirical validation on large-scale real-world data:** Using seven months of traffic and incident records in Tokyo, we demonstrate that CASTANET consistently outperforms state-of-the-art baselines, with particularly large improvements in post-incident and severe-congestion scenarios where accurate prediction matters most for operational decisions.

II. RELATED WORK

A. Traffic Forecasting

Traffic engineering has developed mathematical models to simulate traffic flow by assuming theoretical conditions. Macroscopic models, such as queueing theory [13] and kinematic wave theory [14], attempt to simulate large-scale traffic. Microscopic models, such as car-following models [15], aim to simulate individual vehicles' behavior. However, real-world traffic data often violate the theoretical assumptions, and do not accurately match the simulated results. Recently, data-driven methods have become popular in traffic prediction [16]. Researchers have attempted to capture spatio-temporal dependencies by combining GNNs for the spatial domain, and a recurrent neural network (RNN) for the temporal domain. The models have demonstrated improvements on the traffic prediction [3], [2], [4], [17], but they do not incorporate incident data. Various methods for predicting future incidents from traffic data have been developed and have successfully improved their performance [6] because incidents tend to occur under certain traffic conditions. These results suggest that incidents are not random in space and time. In this context, studies have considered the impact of incidents on traffic [18], [5]. However, using traffic incident data to predict future traffic is out of their focus.

B. Causal Inference

Treatment effect estimation in causal inference is a method for estimating the change in outcomes due to the selection of treatment [19]. For example, this corresponds to estimating how much the symptom of a disease changes when medication is given to a patient or not. This framework introduces the notion of missing counterfactual outcomes and accounts for selection bias when treatments are not assigned at random [7]. Researchers have proposed bias-reduction methods to accurately estimate average treatment effects given a condition [20]. Econometrics and statistics have extended this framework for spatial data analysis [21]. Recently, deep learning methods for estimating CATE (Conditional Average Treatment Effect) have emerged and received considerable attention in machine learning [22]. Their core idea is to remove

the selection bias by learning representations where treatment and control groups are identically distributed [23], [8]. Since removing bias can prevent overfitting to factual outcomes, they have demonstrated the improved accuracy for not only CATE but for predicting factual outcomes. Various methods have been developed for estimating CATE on multivariate series [9], [10], and spatio-temporal data [11], [12]. For example, they attempted to consider the spread of diseases, the behavior of agents, and the evacuation movements of crowds, but traffic data is out of their scope. Existing studies have utilized causal relationships between time-series data and a road network for graph structure learning [24], [25] or introduce causal inference to capture relationships between multi-modal time-series data [26]. However, these models have not considered sudden traffic incidents and their effect on traffic.

III. PROBLEM SETTING

We represent a road network by an adjacency matrix $W \in \{0, 1\}^{N \times N}$, where N is the number of road segments. The element w_{ij} is set to 1 if segments i and j are connected, or 0 otherwise. Let $\mathbf{X}_t^{\text{traffic}} \in \mathbb{R}^{N \times d_{\text{traffic}}}$ denote traffic observations at time step t , where d_{traffic} is the number of traffic variables. $\mathbf{X}_t^{\text{traffic}}$ includes queue length, speed, and flow for each segment. We use auxiliary features $\mathbf{X}_t^{\text{aux}} \in \mathbb{R}^{N \times d_{\text{aux}}}$, such as time-of-day, day-of-week, and static road attributes. We concatenate them as $\mathbf{X}_t = [\mathbf{X}_t^{\text{traffic}}, \mathbf{X}_t^{\text{aux}}] \in \mathbb{R}^{N \times d_{\text{feature}}}$, where $d_{\text{feature}} = d_{\text{traffic}} + d_{\text{aux}}$.

We represent the current incident status by a binary vector $\mathbf{a}_t \in \{0, 1\}^N$, where $a_{i,t} = 1$ if an incident occurs on segment i at time t , or $a_{i,t} = 0$ otherwise. We denote the traffic history over the last T time steps by $\mathcal{X}_t = [\mathbf{X}_{t-T+1}, \dots, \mathbf{X}_t] \in \mathbb{R}^{N \times d_{\text{feature}} \times T}$, and the incident history (excluding the current time step) by $\mathbf{A}_{t-1} = [\mathbf{a}_{t-T}, \dots, \mathbf{a}_{t-1}] \in \{0, 1\}^{N \times T}$. Let $\mathbf{y}_t \in \mathbb{R}^N$ denote the target congestion variable at time t (queue length in our experiments), and define the prediction target for the next T' time steps as $\mathbf{Y}_t = [\mathbf{y}_{t+1}, \dots, \mathbf{y}_{t+T'}] \in \mathbb{R}^{N \times T'}$.

Given a dataset $D = \{(\mathcal{X}_t, \mathbf{A}_{t-1}, \mathbf{a}_t, \mathbf{Y}_t, W)\}_{t=1}^\tau$ with τ time steps, our goal is to learn a predictor f_θ that outputs $\hat{\mathbf{Y}}_t$:

$$\hat{\mathbf{Y}}_t = f_\theta(\mathcal{X}_t, \mathbf{A}_{t-1}, \mathbf{a}_t, W). \quad (1)$$

We estimate parameters $\hat{\theta}$ by minimizing the empirical risk with a loss function $\ell(\cdot, \cdot)$:

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{\tau} \sum_{t=1}^{\tau} \ell(\mathbf{Y}_t, \hat{\mathbf{Y}}_t). \quad (2)$$

IV. CASTANET

As shown in Fig. 2, sudden incidents such as traffic accidents are not uniformly distributed across time and space. To mitigate this issue, we build on the treatment effect estimation framework [27], [7].

A. Causal Viewpoint

We treat the traffic history $\mathcal{X}_t$, incident history $\mathbf{A}_{t-1}$, and road network W as covariates, the current incident status $\mathbf{a}_t$ as a treatment, and the future traffic $\mathbf{Y}_t$ as the outcome. We

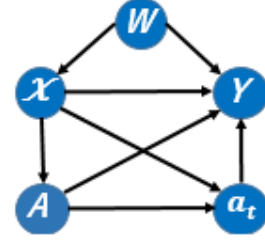


Fig. 3. A causal graph representing the relationship between traffic data and incidents. $\mathcal{X}$, $\mathbf{A}$, $\mathbf{Y}$, $\mathbf{a}_t$, and W denote the traffic history, incident history, outcome, current incident status, and road network, respectively.

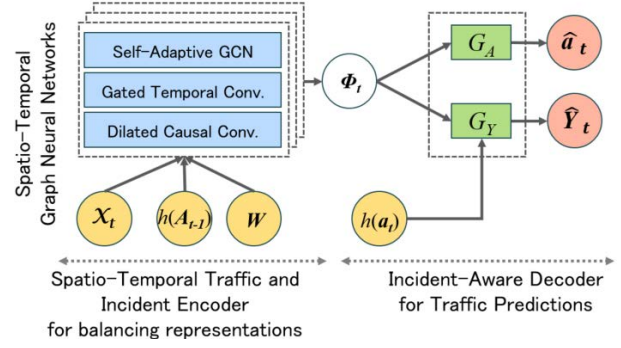


Fig. 4. The overview of CASTANET. Incident embedding h transforms the incident history $\mathbf{A}_{t-1}$ and the current incident status $\mathbf{a}_t$. Spatio-temporal encoder uses STGNNs to extract latent representation Φ_t from traffic data $\mathcal{X}_t$ and embedding $h(\mathbf{A}_{t-1})$ using an adjacency matrix W . Decoder predicts future traffic $\hat{\mathbf{Y}}_t$ and the current incident status $\hat{\mathbf{a}}_t$ with G_Y and G_A .

assume that these variables follow the causal graph shown in Fig. 3. The outcome $\mathbf{Y}_t$ depends on $\mathcal{X}_t$, $\mathbf{A}_{t-1}$, W , and $\mathbf{a}_t$, while the occurrence of incidents is conditioned on recent covariates. For simplicity, we assume that incidents occur independently across road segments. We define the propensity score for $a_{i,t}$ as the conditional probability:

$$P(a_{i,t} = 1 \mid \mathcal{X}_t, \mathbf{A}_{t-1}) = s_{i,t}. \quad (3)$$

Following [9], [10], we assume three basic conditions:

Assumption 1: Consistency. If the realized treatment for segment i at time t is $a_{i,t}$, then the observed outcome equals the corresponding potential outcome under $a_{i,t}$.

Assumption 2: Positivity. The propensity score satisfies $0 < P(a_{i,t} = 1 \mid \mathcal{X}_t, \mathbf{A}_{t-1}) < 1$.

Assumption 3: Sequential strong ignorability. Given the sequence of recent covariates, the current treatment assignment is conditionally independent of future potential outcomes.

Based on this viewpoint, we propose a causality-aware spatio-temporal adversarial network (CASTANET) that incorporates incident information to predict future traffic. CASTANET integrates spatio-temporal graph neural networks (STGNNs) [2] and treatment effect estimation. Figure 4 shows an overview of CASTANET. We extend deep representation learning methods for multivariate time series [9], [10] to incorporate spatial traffic relationships through the road network W .

B. Spatio-Temporal Traffic and Incident Encoder

Incident embedding: We represent the incident status of segment i at time step t by a binary indicator $a_{i,t}$, which may be too limited to capture the relationship between incidents and traffic. We therefore embed the incident status into a learnable vector space [12]:

$$h(a_{i,t}) = \begin{cases} \mathbf{u}^{(0)} & \text{if } a_{i,t} = 0, \\ \mathbf{u}^{(1)} & \text{if } a_{i,t} = 1, \end{cases} \quad (4)$$

where $\mathbf{u}^{(0)}, \mathbf{u}^{(1)} \in \mathbb{R}^{d_{\text{emb}}}$ are learnable d_{emb} -dimensional vectors. We obtain incident embeddings from the historical records $\mathbf{A}_{t-1}$ as $h(\mathbf{A}_{t-1}) = [h(\mathbf{a}_{t-T}), \dots, h(\mathbf{a}_{t-1})] \in \mathbb{R}^{N \times d_{\text{emb}} \times T}$, and from the current incident status $\mathbf{a}_t$ as $h(\mathbf{a}_t) \in \mathbb{R}^{N \times d_{\text{emb}}}$.

Latent spatio-temporal representation: We develop a spatio-temporal encoder to extract temporal and spatial contexts from historical traffic data, incident history, and the road network. Following balancing representation methods [9], an STGNN f_{STGNN} extracts a latent representation $\Phi_t \in \mathbb{R}^{N \times d_{\text{rep}}}$ from $\mathcal{X}_t$, $h(\mathbf{A}_{t-1})$ and $\mathbf{W}$:

$$\Phi_t = f_{\text{STGNN}}(\mathcal{X}_t, h(\mathbf{A}_{t-1}), \mathbf{W}; \theta_R), \quad (5)$$

where θ_R is a set of STGNN parameters. The representation Φ_t is designed to be predictive of the future outcome $\mathbf{Y}_t$ while being less predictive of the current incident status $\mathbf{a}_t$, which helps mitigate bias in incident occurrences.

We employ Graph WaveNet (GWNNet) [2] as the STGNN backbone, as it learns a self-adaptive adjacency matrix and captures dependencies beyond the predefined graph structure, which is particularly beneficial for complex road networks such as Tokyo. We concatenate the traffic history $\mathcal{X}_t$ and incident embeddings $h(\mathbf{A}_{t-1})$ as $[\mathcal{X}_t, h(\mathbf{A}_{t-1})]$ and use it as input to GWNNet. GWNNet consists of K recursive temporal and spatial layers. The temporal layer applies a dilated causal convolution to extract historical features using only data prior to a given time, and a gated temporal convolution to adaptively select features.

The spatial layer employs graph convolution networks (GCNs) [28] with both predefined and learned adjacency matrices. Let $\mathbf{E}_1, \mathbf{E}_2 \in \mathbb{R}^{N \times d_{\text{node}}}$ be learnable node embeddings. The learned adjacency matrix is computed as $\mathbf{P} = \text{Softmax}(\text{ReLU}(\mathbf{E}_1 \mathbf{E}_2^\top))$. By learning $\mathbf{P}$, we mitigate the mismatch between real traffic propagation patterns and the hand-crafted graph $\mathbf{W}$. We denote the self-adaptive GCN as

$$f_{\text{SAGCN}}(\mathbf{H}) = \sum_{k=0}^K (\mathbf{P}_f^k \mathbf{H} \mathbf{S}_{k1} + \mathbf{P}_b^k \mathbf{H} \mathbf{S}_{k2} + \mathbf{P}^k \mathbf{H} \mathbf{S}_{k3}), \quad (6)$$

where $\mathbf{H} \in \mathbb{R}^{N \times d_{\text{rep}}}$ is an input feature matrix, $\mathbf{S}_{k1}, \mathbf{S}_{k2}, \mathbf{S}_{k3} \in \mathbb{R}^{d_{\text{rep}} \times d_{\text{rep}}}$ are learnable parameters and $\mathbf{P}_f, \mathbf{P}_b \in \mathbb{R}^{N \times N}$ are row/column-normalized matrices derived from $\mathbf{W}$.

C. Incident-Aware Decoder

Our decoder uses the representation Φ_t and the current incident embedding $h(\mathbf{a}_t)$ to predict future traffic $\hat{\mathbf{Y}}_t$. We build an outcome predictor network G_Y using a self-adaptive GCN and multilayer perceptrons (MLPs). We first apply Eq. (6) to propagate the current incident information to neighbors:

$$\Psi_t = f_{\text{SAGCN}}([\Phi_t, h(\mathbf{a}_t)]). \quad (7)$$

When the target $\hat{\mathbf{Y}}_t$ is queue length, we employ a traffic engineering-informed layer called the queueing theory (QT) layer [29] and MLPs within G_Y :

$$\hat{\mathbf{V}}_t, \hat{\mathbf{Q}}_t, \hat{\mathbf{C}}_t = f_{\text{MLP1}}([\Phi_t, \Psi_t]), \quad \hat{\mathbf{Y}}_t = f_{\text{QT}}(\hat{\mathbf{V}}_t, \hat{\mathbf{Q}}_t, \hat{\mathbf{C}}_t), \quad (8)$$

where $\hat{\mathbf{V}}_t \in \mathbb{R}^{N \times T'}$, $\hat{\mathbf{Q}}_t \in \mathbb{R}^{N \times T'}$, and $\hat{\mathbf{C}}_t \in \mathbb{R}^{N \times T'}$ are the predicted speed, flow, and correction terms. We denote by θ_Y the parameters in Eq. (7) and Eq. (8).

In addition, we build a treatment classification network G_A to predict the current incident status from Φ_t by employing MLPs:

$$\hat{\mathbf{a}}_t = \sigma(f_{\text{MLP2}}(\Phi_t; \theta_A)), \quad (9)$$

where $\sigma(\cdot)$ is the sigmoid function and θ_A is learnable parameters.

D. Adversarial Representation Learning (ARL)

To mitigate the spatio-temporal selection bias and extreme sparsity of incident records, we incorporate an adversarial objective that discourages the latent representation Φ_t from encoding the current incident status $\mathbf{a}_t$. Following [10], we implement this idea using the counterfactual domain confusion (CDC) loss. This helps (a) improve prediction accuracy for future traffic and (b) reduce reliance on factors that merely make incidents more (or less) likely under specific conditions.

For objective (a), we introduce a loss function $\mathcal{L}_Y(\theta_Y, \theta_R)$ for the outcome predictor G_Y . We use queue length as the main target variable and speed/flow as auxiliary outputs. Following [29], we define the per-time-step losses as

$$\mathcal{L}_{t,q} = \sum_{i=1}^N \sum_{k=1}^{T'} \frac{r_{i,k}}{NT'} |Y_{i,k}^t - \hat{Y}_{i,k}^t|, \quad (10)$$

$$\mathcal{L}_{t,nq} = \sum_{i=1}^N \sum_{k=1}^{T'} \frac{r_{i,k}}{NT'} \frac{|Y_{i,k}^t - \hat{Y}_{i,k}^t|}{L_i}, \quad (11)$$

$$\mathcal{L}_{t,s} = \sum_{i=1}^N \sum_{k=1}^{T'} \frac{1}{NT'} \frac{|V_{i,k}^t - \hat{V}_{i,k}^t|}{V_{i,k}^t}, \quad (12)$$

$$\mathcal{L}_{t,f} = \sum_{i=1}^N \sum_{k=1}^{T'} \frac{1}{NT'} |Q_{i,k}^t - \hat{Q}_{i,k}^t|, \quad (13)$$

where $\mathcal{L}_{t,q}$, $\mathcal{L}_{t,nq}$, $\mathcal{L}_{t,s}$, and $\mathcal{L}_{t,f}$ are the losses for queue length, normalized queue length (segment congestion rate),

speed, and flow, respectively. L_i denotes the length of the i -th road segment. Y^t , V^t and Q^t are the ground truth values for queue length, speed and flow respectively. We define the weight $r_{i,k}$ [29] based on queue length as: $r_{i,k} = r_0$ if $Y_{i,k}^t = 0$, and $r_{i,k} = \sigma\left(\frac{Y_{i,k}^t - \delta_1}{\delta_2}\right)$ otherwise, where r_0 , δ_1 , and δ_2 are hyperparameters. A smaller δ_1 assigns higher weights to moderate queues, whereas a larger δ_1 shifts the emphasis toward severe/long queues; δ_2 controls the sharpness of this transition. In addition, r_0 is the weight for zero-queue samples, introduced to account for the zero-inflated nature of queue-length data (as also noted in [29]) and to prevent the abundant zero cases from dominating training.

We then define $\mathcal{L}_Y(\theta_Y, \theta_R)$ as a weighted sum:

$$\mathcal{L}_Y(\theta_Y, \theta_R) = \sum_{t=1}^{\tau} \left(\lambda_1 \mathcal{L}_{t,q} + \lambda_2 \mathcal{L}_{t,nq} + \lambda_3 \mathcal{L}_{t,s} + \lambda_4 \mathcal{L}_{t,f} \right), \quad (14)$$

where $\lambda_i (i = 1, \dots, 4)$ are hyperparameters for scale adjustment.

To reduce sensitivity to sparse and biased incidents, we employ a counterfactual domain confusion loss [10]. This objective trains the treatment classification network G_A while encouraging the representation Φ_t to be less predictive of the current treatment. We use a weighted cross-entropy loss [30]:

$$\begin{aligned} \mathcal{L}_A(\theta_A, \theta_R) = & -\frac{1}{\tau N} \sum_{t=1}^{\tau} \sum_{i=1}^N \left[\alpha_1 a_{i,t} \log \hat{a}_{i,t} \right. \\ & \left. + \alpha_2 (1 - a_{i,t}) \log (1 - \hat{a}_{i,t}) \right], \end{aligned} \quad (15)$$

where α_1 and α_2 are hyperparameters to handle the rarity of incidents.

We further encourage *unpredictability* by measuring the cross-entropy between outputs of the treatment classification network and a uniform distribution:

$$\mathcal{L}'_A(\theta_A, \theta_R) = -\frac{1}{\tau N} \sum_{t=1}^{\tau} \sum_{i=1}^N \left[\frac{1}{2} \log \hat{a}_{i,t} + \frac{1}{2} \log (1 - \hat{a}_{i,t}) \right]. \quad (16)$$

Based on Eqs. (14), (15), and (16), we train the model by iteratively updating θ_Y , θ_R , and θ_A [10]:

$$(\hat{\theta}_Y, \hat{\theta}_R) = \underset{\theta_Y, \theta_R}{\operatorname{argmin}} \mathcal{L}_Y(\theta_Y, \theta_R) + \lambda_{\text{Conf}} \mathcal{L}'_A(\hat{\theta}_A, \theta_R), \quad (17)$$

$$\hat{\theta}_A = \underset{\theta_A}{\operatorname{argmin}} \mathcal{L}_A(\theta_A, \hat{\theta}_R), \quad (18)$$

where λ_{Conf} is a hyperparameter controlling the strength of domain confusion.

V. EXPERIMENTS

We conducted experiments using real-world data to evaluate the queue length prediction performance of CASTANET. In the experiments, we aim to answer the following four research questions. **RQ1:** Does CASTANET outperform state-of-the-art baseline models in queue length prediction, particularly in

post-incident scenarios? **RQ2:** Does each component of CASTANET contribute to improving the prediction performance? **RQ3:** How does the proposed model actually predict traffic congestion? **RQ4:** How does the proposed model scale when applied to a significantly larger road network?

A. Dataset

We used seven months of traffic and accident data collected by the Tokyo traffic control system from May 1, 2022, to November 30, 2022. In the dataset, only accidents were considered as incidents that affect future traffic. The targeted road network includes $N = 2,032$ road segments in central Tokyo, covering a total of 1,468 km. The traffic data include three variables ($d_{\text{traffic}} = 3$: queue length, speed, and flow for each segment). Although the data were collected at 50 s intervals, they contain significant noise. Therefore, we downsampled them to 5 min intervals and applied smoothing.

The data were processed by the traffic control system based on information from road sensors. The system estimates vehicle density over the road network from densities measured directly by the sensors and flags areas with densities exceeding a predefined threshold as congested. This allows the identification of congested areas within the entire road network, which is then used to calculate queue length. In traffic signal control, it is necessary to consider the number of cars in a queue at each intersection. However, sparsely installed sensors make the accurate estimation of this variable infeasible, so signal control is based on queue lengths [1], [31].

To evaluate robustness, we constructed five datasets using data from the earliest date spanning (a) 3, (b) 4, (c) 5, (d) 6, and (e) 7 months, and independently trained and evaluated models for each. Each dataset was split in a 7:1:2 ratio in chronological order for training, validation, and test. Incident status is denoted by “1” at the time and location of a registered incident; all other data points are denoted by “0.”

The model input includes traffic data from the last $T = 12$ timestamps (1 h), additional features, and current incident status. The output is the queue length for the next $T' = 12$ timestamps (1 h). The additional features consist of the information about the periodicity and heterogeneity of the data, following [29]. We utilized the average of the three traffic variables on each road segment. We also employed the information about each road segment, including the segment length, the number of inflow/outflow segments, the number of lanes and the number of intersections.

B. Metrics

We evaluate the congestion prediction performance of the model using RMSE:

$$\text{RMSE} = \sqrt{\frac{1}{\tau T' N} \sum_{t=1}^{\tau} \sum_{i=1}^N \sum_{t'=1}^{T'} (\hat{Y}_{i,t+t'} - Y_{i,t+t'})^2}. \quad (19)$$

For a test dataset $D = \{(\mathcal{X}_t, \mathbf{A}_{t-1}, \mathbf{a}_t, \mathbf{Y}_t, \mathbf{W})\}_{t=1}^{\tau}$ consisting of observations for τ time steps, we measure the prediction performance after an incident, such as accidents. Let $S =$

TABLE I

RMSE AND RMSE (INCIDENT) FOR ONE-HOUR-AHEAD QUEUE LENGTH PREDICTIONS. THE BEST-PERFORMING SCORES ARE SHOWN IN BOLD, AND THE SECOND-BEST SCORES ARE UNDERLINED. PARENTHESES INDICATE STANDARD DEVIATION. ALL MODELS RECEIVE THE SAME INCIDENT INDICATORS AS ADDITIONAL CHANNELS.

Model	All		Non-zero		Top 10%		Top 5%	
	RMSE[m]	RMSE (Incident)[m]	RMSE[m]	RMSE (Incident)[m]	RMSE[m]	RMSE (Incident)[m]	RMSE[m]	RMSE (Incident)[m]
HA	47.80 _(2.18)	205.76 _(42.45)	90.23 _(3.38)	239.63 _(47.74)	243.49 _(10.05)	417.13 _(68.78)	308.90 _(13.02)	474.00 _(85.49)
Lasso	177.70 _(15.07)	247.09 _(46.16)	249.15 _(17.08)	257.86 _(43.86)	325.30 _(21.36)	387.19 _(62.64)	373.92 _(23.85)	436.77 _(80.62)
Ridge	171.62 _(15.43)	246.51 _(46.20)	243.45 _(20.12)	331.64 _(11.09)	317.42 _(18.54)	407.45 _(52.52)	364.43 _(17.24)	434.84 _(79.94)
XGBoost	64.69 _(6.04)	224.08 _(16.30)	118.66 _(9.24)	260.52 _(15.74)	266.66 _(12.26)	351.46 _(64.31)	328.29 _(12.37)	392.32 _(72.39)
DCRNN	45.25 _(2.82)	181.39 _(42.98)	88.27 _(3.53)	210.98 _(47.75)	236.81 _(5.52)	331.81 _(64.15)	295.22 _(6.84)	368.97 _(77.92)
GWNNet	44.15 _(2.67)	180.06 _(17.79)	86.18 _(3.30)	209.93 _(19.11)	230.12 _(5.79)	316.01 _(31.32)	286.51 _(7.48)	353.23 _(39.34)
AGCRN	45.10 _(2.95)	178.76 _(40.74)	88.05 _(3.86)	207.90 _(44.19)	237.98 _(7.91)	333.11 _(64.76)	297.38 _(10.45)	365.52 _(79.39)
MegaCRN	44.99 _(2.85)	196.21 _(28.59)	87.75 _(3.55)	228.43 _(29.45)	235.60 _(4.75)	367.46 _(51.53)	294.13 _(5.59)	386.57 _(44.87)
QTNet	<u>43.57</u> _(2.45)	192.05 _(44.50)	84.23 _(2.84)	223.24 _(49.91)	207.78 _(3.05)	315.91 _(89.22)	259.64 _(4.07)	347.96 _(108.05)
PDG2Seq	46.30 _(1.27)	184.59 _(21.89)	90.58 _(3.43)	215.47 _(26.67)	241.29 _(12.48)	287.72 _(65.21)	300.95 _(14.30)	316.04 _(85.18)
STAEFormer	45.31 _(2.69)	182.11 _(25.26)	88.29 _(3.30)	212.28 _(27.52)	235.56 _(7.66)	316.02 _(71.40)	293.72 _(10.53)	342.91 _(91.27)
CASTANET:NoARL1	43.60 _(2.48)	<u>178.36</u> _(34.16)	<u>83.92</u> _(3.02)	<u>207.11</u> _(36.51)	<u>204.41</u> _(4.63)	301.61 _(65.76)	254.11 _(4.92)	334.06 _(74.09)
CASTANET:NoARL2	48.41 _(1.44)	192.79 _(28.69)	94.04 _(4.87)	223.15 _(28.13)	237.99 _(26.32)	350.85 _(49.88)	294.45 _(32.05)	389.07 _(51.94)
CASTANET(Proposed Model)	41.83 _(2.41)	160.34 _(26.09)	80.94 _(2.87)	186.40 _(28.04)	200.98 _(3.99)	249.09 _(46.93)	250.69 _(4.35)	270.05 _(51.13)

$\{(i, t) \mid 1 \leq i \leq N, 1 \leq t \leq \tau\}$. We define the incident subset as $D' = \{(i, t) \in S \mid a_{i,t} = 1\}$, and report RMSE on D' as RMSE (Incident).

Since the queue length data are zero-inflated, a model that predicts zero for all data points may achieve artificially high performance. To address this, we divide the test dataset into four subsets and evaluate the performance for each:

- *All*: All data points.
- *Non-zero*: Data points where queue length is not zero.
- *Top 10%*: The top 10% of data points in terms of queue length among the non-zero data points.
- *Top 5%*: The top 5% of data points in terms of queue length among the non-zero data points.

C. Baseline models

We employ four conventional models as baselines, including historical averages (HA), Lasso, Ridge, and XGBoost. We additionally use seven state-of-the-art traffic prediction models: DCRNN [3], GWNNet [2], AGCRN [4], MegaCRN [17], QTNet [29], PDG2Seq [32] and STAEFormer [33]. All information is concatenated along the channel direction and fed into each model. The output is the queue length for the next hour. Based on [29], for models other than QTNet, the loss function uses only the terms related to queue length from Eq. (10) and Eq. (11), excluding the weight coefficient $r_{i,k}$. We also evaluate two variants of CASTANET without adversarial representation learning ($\lambda_{\text{Conf}} = 0$): (i) NoARL1, which applies inverse-propensity weighting to Eq. (14) to mitigate treatment-selection bias, and (ii) NoARL2, which reweights Eq. (14) by the inverse empirical incident frequency to alleviate incident sparsity.

D. Experimental Settings

Hyperparameters are tuned to improve the accuracy of queue length on the validation dataset. We set $d_{\text{emb}} = 16$,

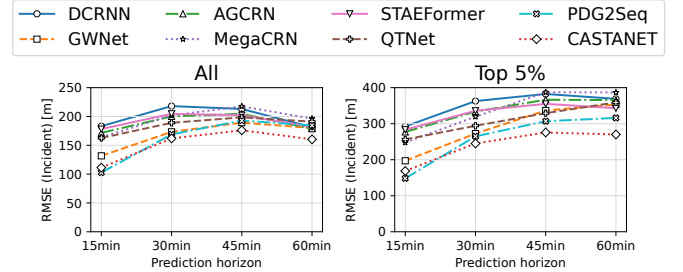


Fig. 5. RMSE (Incident) for one-hour-ahead predictions.

$d_{\text{rep}} = 40$, and $d_{\text{node}} = 10$. The parameters for the loss function were set to $\lambda_1 = 4.0$, $\lambda_2 = 0.02$, $\lambda_3 = 0.05$, $\lambda_4 = 0.08$, $\delta_1 = 100$, $\delta_2 = 100$, $\lambda_{\text{Conf}} = 0.1$, $\alpha_1 = 10$, and $\alpha_2 = 0.01$. The loss weights were set so that the magnitudes of the queue-length, normalized queue-length, speed, and flow loss terms are comparable on the validation set, preventing any single term from dominating optimization. Additionally, curriculum learning was applied to efficiently train the outcome predictor network using the QT-layer, following the setup described in [29]. For model optimization, we used Adam, with the learning rate starting at 0.001 and decaying by a factor of 0.97 each epoch. Training was conducted for 200 epochs, with early stopping applied if the validation loss did not improve for 10 consecutive epochs. All experiments were conducted using Python 3.8.10 and PyTorch 1.13.1 on a server with an NVIDIA A100 GPU. We needed propensity scores $s_{i,t}$ for CASTANET:NoARL1. We built a neural network to predict propensity scores given $\mathcal{X}_t$ and $\mathcal{A}_{t-1}$. Trained to minimize Equation (15), this model achieved an AUC of 0.84 for current incident status prediction, indicating the validity of the estimated propensity scores.

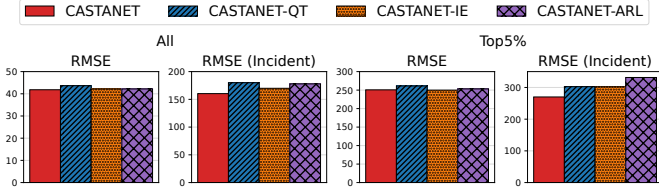


Fig. 6. Ablation study on RMSE and RMSE (Incident).

E. Results

To address **RQ1**, we evaluated the average and standard deviation of RMSE across the datasets (a)–(e). We show the results in Table I. We confirmed that CASTANET improved queue length predictions regardless of whether an incident occurred. In the All setting, CASTANET reduces RMSE by 4.0% compared with the best baseline (QTNet), and reduces RMSE (Incident) by 10.1% compared with the best competing method (CASTANET:NoARL1). Under severe congestion (Top 5%), CASTANET outperformed the state-of-the-art models by a wide margin of 14.55% in RMSE (Incident), which demonstrates the effectiveness of our proposed architecture for utilizing incident data. CASTANET also outperformed NoARL1 and NoARL2, indicating that ARL provides additional benefits beyond inverse-propensity weighting and incident-frequency reweighting in our setting. NoARL2 degrades performance compared to CASTANET because aggressive weighting of the rare treated class increases variance. Unlike sample reweighting, ARL explicitly enforces treatment-invariant representations, which helps the model utilize incident information. HA achieves comparable RMSE to STGNN-based baselines, but performs poorly on RMSE (Incident) particularly in Top 10% and Top 5%. This result indicates that most traffic patterns are periodic. However, sudden incidents cause non-periodic congestion, which is difficult for HA to predict. Despite this difficulty, CASTANET can predict both periodic and non-periodic congestion. PDG2Seq utilizes dynamic graph convolution, which is particularly effective when accidents occur, resulting in the second-best performance. Figure 5 shows the progression of RMSE (Incident) up to one-hour-ahead for CASTANET and baseline STGNN-based models. The performance gain of CASTANET persists over longer horizons, suggesting that it captures not only the immediate post-incident impact but also the sustained congestion dynamics post-incident. PDG2Seq outperforms CASTANET in the 15-min-ahead prediction setting because PDG2Seq utilizes dynamic graph convolution.

F. Ablation Study

To address **RQ2**, we conducted an ablation study comparing CASTANET with its variants; **CASTANET-IE**: CASTANET without the incident embedding, **CASTANET-QT**: CASTANET without the QT-layer, and **CASTANET-ARL**: CASTANET without ARL. Figure 6 presents the one-hour-ahead RMSE and RMSE (Incident) averaged over the datasets (a)–(e) for CASTANET and its variants. Compared to CAS-

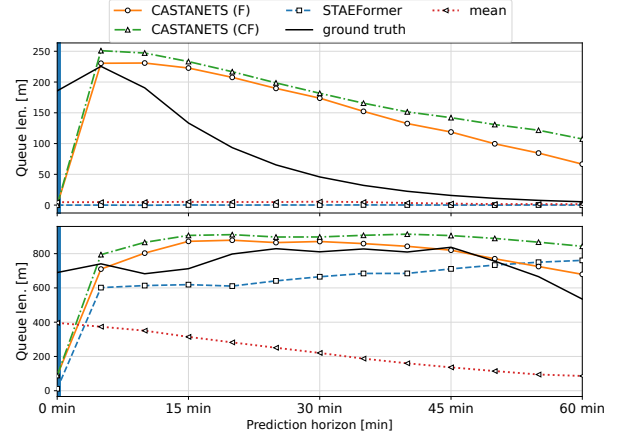


Fig. 7. Every five-minute factual and counterfactual predictions by CASTANET. The incident time is indicated by the blue vertical line at 0 min on the x-axis.

TANET, all variants yielded a performance decline, indicating the contribution of each component to improving prediction accuracy. Of these components, ARL contributed the most to the improvement in performance, particularly with regard to RMSE (Incident). Because ARL mitigates bias and overfitting in the incident distribution of the training data, we expect it to generalize better by providing unbiased predictions on the test data. The QT-layer also enhances prediction accuracy due to its traffic engineering constraints. Incident embedding consistently improves performance across all subsets, indicating that it effectively conveys incident information to the model.

G. Qualitative Evaluation

To address **RQ3**, we visualize the predicted outputs of CASTANET. Since CASTANET is a model based on causal inference methods, we visualize the factual prediction (CASTANET (F)) where an incident is provided as an input, and the counterfactual prediction (CASTANET (CF)), in which we set the current incident indicator to zero. We show every five-minute predictions up to one-hour-ahead in Figure 7. For simplicity, STAEFormer is shown here for a comparison. In both cases, CASTANET (F), which uses factual incident data, provides more accurate predictions than CASTANET (CF), which uses counterfactual incident data. Moreover, the effect of the incident indicator varies across cases, suggesting that CASTANET adjusts its response to incidents according to the traffic context rather than applying a fixed rule.

H. Computation Time

To address **RQ4**, we measure the computational time of CASTANET and STAEFormer on a single NVIDIA A100 GPU. For the Tokyo dataset, CASTANET requires 0.02 seconds to generate multi-horizon predictions from 5 to 60 minutes ahead and training for 84 epochs took 641 minutes. STAEFormer requires 0.05 seconds to generate predictions and training for 20 epochs took 656 minutes. This inference and training time are comparable to that of the STGNN-based

baseline. In our current setting, CASTANET uses approximately 300 MB of GPU memory. We believe that CASTANET can be applied to a significantly larger road network or city-wide traffic control system.

VI. CONCLUSION

In this paper, we proposed CASTANET, which integrates an STGNN backbone with balanced representation learning from the treatment effect estimation framework to predict traffic congestion under sudden incidents. In future work, we plan to incorporate incidents beyond accidents, to examine the sensitivity to violations of the underlying causal assumptions and to introduce an off-policy evaluation metric.

ACKNOWLEDGMENTS

The authors would like to thank the Traffic Facilities and Control Division, Traffic Bureau, Tokyo Metropolitan Police Department for providing the traffic data. This work was supported by JST FOREST (JPMJFR232S), JST BOOST (JPMJBY25D0), and JSPS KAKENHI (26K02984).

REFERENCES

- [1] H. Sakakibara, T. Usami, S. Itakura, and T. Tajima, “Moderato (management by origin-destination related adaptation for traffic optimization),” in *Proceedings 1999 IEEE/IEEEJ/ISAI International Conference on Intelligent Transportation Systems*, 1999.
- [2] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, “Graph wavenet for deep spatial-temporal graph modeling,” in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 1907–1913.
- [3] Y. Li, R. Yu, C. Shahabi, and Y. Liu, “Diffusion convolutional recurrent neural network: Data-driven traffic forecasting,” in *The 6th International Conference on Learning Representations*, 2018.
- [4] L. Bai, L. Yao, C. Li, X. Wang, and C. Wang, “Adaptive graph convolutional recurrent network for traffic forecasting,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 17804–17815.
- [5] Y. Lin and R. Li, “Real-time traffic accidents post-impact prediction: Based on crowdsourcing data,” *Accident Analysis & Prevention*, vol. 145, p. 105696, 2020.
- [6] S. Moosavi, M. H. Samavatian, S. Parthasarathy, R. Teodorescu, and R. Ramnath, “Accident risk prediction based on heterogeneous sparse data: New dataset and insights,” in *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2019, p. 33–42.
- [7] D. B. Rubin, “Causal inference using potential outcomes,” *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 322–331, 2005.
- [8] U. Shalit, F. D. Johansson, and D. Sontag, “Estimating individual treatment effect: generalization bounds and algorithms,” in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, 2017, pp. 3076–3085.
- [9] I. Bica, A. M. Alaa, J. Jordon, and M. van der Schaar, “Estimating counterfactual treatment outcomes over time through adversarially balanced representations,” *International Conference on Learning Representations*, 2020.
- [10] V. Melnychuk, D. Frauen, and S. Feuerriegel, “Causal transformer for estimating counterfactual outcomes,” in *Proceedings of the 39th International Conference on Machine Learning*, vol. 162, 2022, pp. 15293–15329.
- [11] J. Ma, Y. Dong, Z. Huang, D. Mietchen, and J. Li, “Assessing the causal impact of covid-19 related policies on outbreak dynamics: A case study in the us,” in *Proceedings of the ACM Web Conference*, 2022.
- [12] K. Takeuchi, R. Nishida, H. Kashima, and M. Onishi, “Causal effect estimation on hierarchical spatial graph data,” in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, p. 2145–2154.
- [13] E. Cascetta, *Transportation Systems Engineering: Theory and Methods*. Springer New York, 2013.
- [14] C. F. Daganzo and J. A. Laval, “Moving bottlenecks: A numerical method that converges in flows,” Institute of Transportation Studies at UC Berkeley, Tech. Rep., 2003.
- [15] J. J. Olstam and A. Tapani, *Comparison of Car-floating models*. Swedish National Road and Transport Research Institute (VTI), 2004.
- [16] X. Yin, G. Wu, J. Wei, Y. Shen, H. Qi, and B. Yin, “Deep learning on traffic prediction: Methods, analysis, and future directions,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 4927–4943, 2021.
- [17] R. Jiang, Z. Wang, J. Yong, P. Jeph, Q. Chen, Y. Kobayashi, X. Song, S. Fukushima, and T. Suzumura, “Spatio-temporal meta-graph learning for traffic forecasting,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 7, pp. 8078–8086, 2023.
- [18] R. Yu, Y. Li, C. Shahabi, U. Demiryurek, and Y. Liu, “Deep learning: A generic approach for extreme condition traffic forecasting,” in *Proceedings of the 2017 SIAM International Conference on Data Mining (SDM)*, 2017.
- [19] D. B. Rubin, “The use of propensity scores in applied bayesian inference,” in *Bayesian Statistics 2*, 1985, pp. 463–472.
- [20] A. Curth and M. Van der Schaar, “Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms,” in *International Conference on Artificial Intelligence and Statistics*, 2021.
- [21] K. Akbari, S. Winter, and M. Tomko, “Spatial causality: A systematic review on spatial causal inference,” *Geographical Analysis*, vol. 55, no. 1, pp. 56–89, 2023.
- [22] I. Bica, A. M. Alaa, C. Lambert, and M. van der Schaar, “From real-world patient data to individualized treatment effects using machine learning: Current and future methods to address underlying challenges,” *Clinical Pharmacology & Therapeutics*, vol. 109, no. 1, pp. 87–100, 2021.
- [23] F. Johansson, U. Shalit, and D. Sontag, “Learning representations for counterfactual inference,” in *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48, 2016, pp. 3020–3029.
- [24] L. Zhang, K. Fu, T. Ji, and C.-T. Lu, “Granger causal inference for interpretable traffic prediction,” in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, 2022, pp. 1645–1651.
- [25] J. Lin, Z. Li, Z. Li, L. Bai, R. Zhao, and C. Zhang, “Dynamic causal graph convolutional network for traffic prediction,” *arXiv preprint arXiv:2306.07019*, 2023.
- [26] Y. Zhao, P. Deng, J. Liu, X. Jia, and M. Wang, “Causal conditional hidden markov model for multimodal traffic prediction,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 4, pp. 4929–4936, 2023.
- [27] J. Robins and M. Hernán, “Estimation of the causal effects of time-varying exposure,” in *Longitudinal Data Analysis*, G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs, Eds. Boca Raton, FL: Chapman & Hall/CRC, 2008, pp. 553–599.
- [28] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in *The 5th International Conference on Learning Representations*, 2017.
- [29] R. Shirakami, T. Kitahara, K. Takeuchi, and H. Kashima, “Qtmet: Theory-based queue length prediction for urban traffic,” in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, p. 4832–4841.
- [30] H. He and E. A. Garcia, “Learning from imbalanced data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.
- [31] S. Miyata, M. Noda, and T. Usami, “Stream (strategic realtime control for megalopolis-traffic) advanced traffic control system of tokyo,” *The Second World Congress on ITS '95 YOKOHAMA*, pp. 289–297, 1995.
- [32] J. Fan, W. Weng, Q. Chen, H. Wu, and J. Wu, “PDG2Seq: Periodic Dynamic Graph to Sequence Model for Traffic Flow Prediction,” *Neural Networks*, vol. 183, p. 106941, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0893608024008700>
- [33] H. Liu, Z. Dong, R. Jiang, J. Deng, J. Deng, Q. Chen, and X. Song, “Spatio-temporal adaptive embedding makes vanilla transformer sota for traffic forecasting,” in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023, p. 4125–4129.