Cross-Mode Knowledge Adaptation for Bike Sharing Demand Prediction Using Domain-Adversarial Graph Neural Networks

Yuebing Liang, Guan Huang, and Zhan Zhao

Abstract—For bike sharing systems, demand prediction is crucial to ensure the timely re-balancing of available bikes according to predicted demand. Existing methods for bike sharing demand prediction are mostly based on its own historical demand variation, essentially regarding it as a closed system and neglecting the interaction between different transportation modes. This is particularly important for bike sharing because it is often used to complement travel through other modes (e.g., public transit). Despite some recent progress, no existing method is capable of leveraging spatiotemporal information from multiple modes and explicitly considers the distribution discrepancy between them, which can easily lead to negative transfer. To address these challenges, this study proposes a domain-adversarial multi-relational graph neural network (DA-MRGNN) for bike sharing demand prediction with multimodal historical data as input. A spatiotemporal adversarial adaptation network is introduced to extract shareable features from demand patterns of different modes. To capture correlations between spatial units across modes, we adapt a multi-relational graph neural network (MRGNN) considering both geographical proximity and mobility pattern similarity. Extensive experiments are conducted using real-world bike sharing, subway and ride-hailing data from New York City. The results demonstrate the superior performance of our proposed approach compared to existing methods and the effectiveness of different model components.

Index Terms—Bike sharing, demand prediction, inter-modal relationships, graph neural networks, adversarial learning.

I. INTRODUCTION

Bike sharing is an emerging sustainable, convenient and generally affordable mode of transportation. Over the past decade, it has been widely deployed in many cities around the world as an effective solution to traffic congestion and last-mile problems. Due to its positive impact on the environment and public health, bike sharing systems (BSSs) are being promoted to play an increasingly important role in urban transportation systems. However, the efficient operation of BSSs is challenged by fluctuations in spatial and temporal demand patterns, resulting in inefficient bike repositioning and high operating costs for bike rebalancing [1]. Accurate short-term demand prediction at high spatial resolution is the basis to support dynamic rebalancing of available bikes and ensure the user experience of bike sharing service.

Numerous studies have been conducted for the problem of bike sharing demand prediction. Early attempts use regression or machine learning models to solve this problem, which suffer from relatively low accuracy. Recently, more research interests have shifted to deep learning methods, due to their demonstrated effectiveness in extracting the complex knowledge hidden in large-scale mobility data [1], [2]. In particular, Graph Neural Networks (GNNs) have been employed in the demand prediction problem and achieved state-of-the-art performance [3], [4]. Despite the success of these methods, they regard bike sharing as a closed system and neglect the potential rich information of the interaction between BSS and other transportation modes. This is especially important to consider for bike sharing, because it is mainly used for short-distance trips or the first-mile/last-mile portion of longer trips. In practice, BSSs are often designed as feeders to public transit systems or support multimodal transportation connections [5]. As a result, the demand for bike sharing will inevitably be influenced by other transportation modes, which should be considered in demand prediction. Incorporating demand information across modes can also help mitigate the data sparsity problem commonly seen in BSS, since bike sharing is rarely one of the primary travel modes in cities.

To incorporate inter-modal relationships, several recent studies have investigated the joint demand prediction of multimodal transportation systems using multi-task learning frameworks [6], [7], [8]. These methods treat different modes equally and might not be effective enough when we take interest in a particular mode. Another group of research focused on the demand prediction of a target mode by adapting the learned knowledge from other modes [5], [9], [10]. This study belongs to the second category, in which we aim to enhance the prediction performance of BSS with the help of other modes. This is particularly important for bike sharing because it is mainly used for short-distance trips or the first-mile/last-mile portion of longer trips. In practice, BSSs are often designed as feeders to public transit systems or support multimodal transportation connections [5]. As a result, the demand for bike sharing will inevitably be influenced by other transportation modes, which should be considered in demand prediction. Incorporating demand information across modes can also help mitigate the data sparsity problem commonly seen in BSS, since bike sharing is rarely one of the primary travel modes in cities.

To incorporate inter-modal relationships, several recent studies have investigated the joint demand prediction of multimodal transportation systems using multi-task learning frameworks [6], [7], [8]. These methods treat different modes equally and might not be effective enough when we take interest in a particular mode. Another group of research focused on the demand prediction of a target mode by adapting the learned knowledge from other modes [5], [9], [10]. This study belongs to the second category, in which we aim to enhance the prediction performance of BSS with the help of other modes.
of cross-mode demand information. Despite existing relevant works, there still exist several challenges to be resolved.

The first challenge is how to fuse spatiotemporal information from stations or zones of multiple modes for demand prediction of a target mode. To utilize multimodal demand information, existing methods either aggregate the usage of other modes as additional attributes of BSS stations for model input [5], [9], [11], or capture mode-specific temporal dependencies with recurrent-based models and transfer the learned temporal features across modes with knowledge adaptation techniques [10], [12]. They do not consider the spatial dependencies between spatial units of different modes and are not capable of leveraging spatiotemporal knowledge hidden in the multimodal system. Although recent research has introduced graph neural networks to capture cross-mode spatial relationships [8], [13], they are designed for jointly predicting the demand of multiple modes and do not directly consider issues related to domain adaptation from one mode to another, such as the demand distribution discrepancy between modes.

The second challenge is how to handle the distribution discrepancy between the demand patterns of different modes. For instance, our preliminary analysis shows that there is a notable disparity between the demand patterns of bike sharing, subway and ride-hailing in Manhattan: generally, bike sharing usage is more active in the daytime, while the demand of subway is more concentrated during rush hours and ride-hailing is more busy at night (see Section V-A for more details). This is not uncommon in multimodal transportation systems, as people use different modes for various purposes. Directly leveraging multimodal data without reducing the effect of distribution discrepancy can easily lead to negative transfer. This has been confirmed in previous research [9], which showed that simply incorporating subway demand data as model input could have negative effect on bike demand prediction.

To address these issues, in this research, we propose a domain-adversarial multi-relational graph neural network (DA-MRGN) for bike sharing demand prediction based on cross-mode knowledge adaptation. We define multiple intra- and inter-modal graphs to encode the spatial dependencies between stations or zones from different modes. A spatiotemporal adversarial adaptation network is introduced to learn mode-transferable temporal and spatial features and a multi-relational graph neural network [13] is used to enable knowledge adaptation across modes with heterogeneous spatial units. Experiments are conducted using the Citi Bike data from New York City (NYC), with the subway and ride-hailing data as additional inter-modal demand inputs. Because of data constraint, we will focus on station-based BSS in this study, but the methodology should be generalizable to stationless (or dockless) BSS, as well as other similar mobility systems. The specific contributions of this research are as follows: noitemsep

- We propose a graph-based domain adaptation approach to bike sharing demand prediction based on cross-mode relationships learned from historical demand data. With an adapted multi-relational graph neural network, the proposed approach can effectively fuse spatiotemporal information from stations or zones of different modes.
- To reduce distribution discrepancy between modes and improve transfer learning performance, we introduce a spatiotemporal adversarial adaptation network that learns transferable features from auxiliary modes to BSS stations with adversarial learning.
- Extensive experiments are conducted based on real-world datasets from NYC, and the results demonstrate the effectiveness of our proposed model architecture and all the introduced components.

This study extends a preliminary version of the proposed model [14] in the following aspects: (1) with the spatiotemporal adversarial adaptation network to promote positive knowledge adaptation across modes, the performance of our proposed model is improved (see Section IV-A); (2) this work provides more comprehensive experimental results to evaluate how cross-mode information and adversarial learning impact the prediction performance (in Section VI-B and Section VI-C). We also demonstrate the interpretability of our model regarding the spatial dependencies between stations and zones in the multimodal network in Section VI-D.

II. RELATED WORK

In this section, we first review existing methods regarding bike sharing demand prediction, and then provide a brief summary of domain adaptation, which is relevant to our methodology.

A. Bike Sharing Demand Prediction

Traditional methods for bike sharing demand prediction mainly focus on finding the relationships between bike demand and its historical demand through regression models, such as auto-regressive integrated moving average (ARIMA) [15] and linear regression (LR) [16]. Later research uses other more advanced machine learning models. Feng et al. [17] introduced a hierarchical demand prediction model based on gradient boosting regression tree. Guidon et al. [18] estimated the demand of bike sharing services using spatial regression models and random forest. These methods may not be adequate to capture the complex relationships between input variables and bike demand, and thus are not precise enough.

In recent years, extensive studies have presented the power of deep learning models to capture the nonlinear and complex relationships for demand prediction tasks. Xu et al. [1] adopted a long-short-term memory neural network (LSTM) with exogenous factors as additional input. A novel recurrent network was introduced in [2], which explicitly considers multiple historical time steps at each time step. To incorporate spatial information, previous research employed convolutional neural networks (CNNs) with recurrent networks [19], which, however, require aggregating demand data to a grid system. Graph neural networks (GNNs) exempt the requirement of artificial segmentation, which is more useful for BSS operations [20]. Lin et al. [21] proposed a graph convolutional network (GCN)-based approach along with recurrent networks for station-level bike demand prediction. A purely convolutional model was introduced in [4] with GCNs and TCNs to...
model spatial and temporal dependencies respectively. To capture more complex correlations hidden in the transportation network, recent studies used attention-based graph learning approaches to capture dynamic pairwise correlations between stations [22]. A multi-graph convolutional neural network was proposed in [3], using multiple graphs to encode different types of spatial dependencies. Wu et al. [23] proposed to learn spatial dependencies directly from traffic data, with an adaptive adjacency matrix learned through node embedding. However, these methods are mode-specific and do not consider inter-modal relationships.

To leverage the cross-mode demand information, increasing attention has been paid to the co-prediction of multiple transport modes. In [6] and [24], the demands for taxis and bike sharing are aggregated to a grid system to enable shareable feature learning, before co-predicted using a convolutional recurrent network. For demand prediction of general multimodal systems with diverse spatial units, Xu et al. [8] introduced a co-modal Graph Attention Network to capture the interactions between different modes. A multi-relational graph neural network (MRGNN) was developed in our previous work [13] for the joint demand prediction of subway and ride-hailing. These methods pre-assume that different transport modes have similar demand patterns and might not be effective when the distribution of different modes are significantly different. In addition, they treat different transport modes equally and may not work well when we take specific interest in a target mode.

Though limited, several recent works have exploited cross-mode demand information for bike sharing demand prediction. Zhang et al. [5] developed an LSTM-based model with the historical demand of bus and subway as additional input. A graph learning approach was introduced in [9] to enhance the prediction accuracy of BSS demand during peak hours with public transit usage information. Lv et al. [11] focused on the demand prediction of BSS stations around subway stations by extracting subway-related features. These methods process the historical flow of adjacent public transit stations as additional attributes of BSS stations, and do not model the relationships between stations or zones from different modes directly. To capture temporal dependencies across modes, a memory-augmented recurrent model was proposed in [10] for the demand prediction of a station-sparse mode by taking advantage of the demand information of a station-intensive mode. Hua et al. [12] enhanced the prediction performance of bike sharing with a pre-trained LSTM model for public transit using transfer learning strategies. These methods only consider temporal correlations across modes, while a model that can effectively transfer knowledge across modes considering the effect of spatial dependencies between heterogeneous spatial units is needed.

### B. Domain Adaptation

Domain adaptation is a sub-field within transfer learning that aims to reduce the effect of distribution discrepancy when transferring knowledge across domains [25]. To learn transferable representations, previous works focus on minimizing the distribution distance between domains [26]. Tzeng et al. [27] introduced a domain confusion loss based on Maximum Mean Discrepancy (MMD). A Deep Adaptation Network was developed in [28] which reduces the multi-kernel MMD of embedded representations from different domains in a reproducing kernel Hilbert space. Recently, inspired by the idea of adversarial learning, domain adversarial neural networks (DANN) was introduced in [29], which learns domain-invariant feature representations that cannot be distinguished by a domain discriminator. Later research demonstrates the effectiveness of domain adversarial learning for multiple transfer learning tasks in the field of computer vision and natural language processing, such as image classification, face recognition and language understanding [25].

Several recent studies have introduced the idea of adversarial learning to cross-domain human mobility prediction. Tang et al. [30] proposed an adversarial spatiotemporal network for short-term traffic forecasting across cities. A spatial adversarial adaptation module was designed in [31] to capture transferable spatial correlations between source and target cities. Wang et al. [7] developed a multi-task learning framework for crowd-level and OD-level flow prediction, with an adversarial loss for shared feature extraction across tasks. These methods aimed at transferring knowledge across cities or tasks, which are different from this paper, where we use domain adversarial networks to mitigate the distribution discrepancy between multiple transport modes.

### III. Problem Statement

In this section, we introduce some notations in this research and then formulate our problem.

**Definition 1 (Demand Sequence):** Consider a transport mode $m$ with $N_m$ nodes (i.e. stations/service zones). For each node $i = 1, 2, \ldots, N_m$, its inflow and outflow demand at time step $t$ is denoted as $x_{m,i}^t \in \mathbb{R}^2$. Next, we represent the demand of all the nodes from mode $m$ at time step $t$ as $X_m^t = \{x_{m,0}^t, x_{m,1}^t, \ldots, x_{m,N_m}^t\} \in \mathbb{R}^{N_m \times 2}$. Finally, the demand sequence of mode $m$ over time steps $T$ is denoted as $X_m^{T-1} = \{X_m^{t-T}, \ldots, X_m^{1}, X_m^{0}\}$.

**Problem (Bike Sharing Demand Prediction):** This research aims to predict the station-level bike sharing demand given historical demand of BSS as well as other modes. Formally, given the historical demand of bike sharing, denoted as $X_b^{t-\tau_2}$, and that of auxiliary modes, i.e., subway and ride-hailing in our case, denoted as $X_s^{t-\tau_3}$ and $X_r^{t-\tau_3}$, the goal is to predict bike sharing demand $X_b^{t+1}$ at the next time step:

$$X_b^{t+1} = F(X_b^{t-\tau_2}, X_s^{t-\tau_3}, X_r^{t-\tau_3}),$$

where $F(\cdot)$ is the prediction function to be learned by our proposed model. This formulation can be easily adapted to other demand prediction problems with multimodal historical demand as input.

### IV. Methodology

This section presents a domain-adversarial multi-relational graph neural network (DA-MRGGNN) for station-level bike sharing demand prediction by taking advantage of historical demand sequences of subway and ride-hailing systems.
Figure 1 shows the overall architecture of our proposed model. It is composed of stacked adversarial spatiotemporal adaptation networks (STAANs) for multimodal knowledge extraction. STAAN consists of a temporal gated convolution network (TCN) and an intra-modal graph convolution network for each mode to extract mode-specific spatiotemporal features. To address the distribution discrepancy across different modes, a domain adversarial learning technique is applied to learn shareable features across modes. Based on learned representations from STAANs, a multi-relational graph neural network [13] is used to aggregate spatiotemporal features from heterogeneous spatial units of different modes, followed by a prediction layer to generate bike sharing demand prediction. Details of each module are introduced below.

A. Spatiotemporal Adversarial Adaptation Network

To capture spatiotemporal correlations within the historical flow of multiple modes, a naive approach would be to apply a separate spatiotemporal modeling component for each mode. The learned features will then be fused in an MRGNN layer which will be introduced later. However, our preliminary analysis shows that there is a notable disparity between the demand patterns of different modes, and neglecting the effect of cross-mode pattern discrepancy might lead to negative transfer (see Section V-A for more details). To address this issue, we design a spatiotemporal adversarial adaptation network (STAAN) by taking advantage of adversarial learning [29]. The idea of adversarial learning is related to generative adversarial networks (GANs) [32], which typically comprise a generator and a discriminator. The discriminator is used to distinguish the real data distribution and the generated data distribution from the generator. The generator, on the other hand, aims to produce outputs that can confuse the discriminator. In our case, we use a spatiotemporal feature extractor for each mode as the generator to extract features from the demand series of different modes and a domain discriminator to distinguish the learned features from different modes. The feature extractors are trained to confuse the domain discriminator so that they learn mode-transferable feature representations.

1) Spatiotemporal Feature Extractor: Previous mobility prediction approaches typically used RNNs or CNNs to extract temporal features. In this research, we employ a temporal gated convolution network (TCN) proposed by [4] to extract temporal features for each mode, due to its fast training time and simple structures. Briefly, given an input sequence, it uses a 1-D causal convolution to capture the relationships between each time step and its neighborhoods, along with an output gate to control the ratio of information that passes through layers. In our case, a separate TCN layer is applied to each mode to capture mode-specific temporal information, which processes all nodes in the mode in parallel. Mathematically, for a node $i$ in mode $m$, the input of the TCN layer is a sequence of feature vectors denoted as $h_{m,i} \in \mathbb{R}^{K \times c}$, where $K$ is the length of the sequence and $c$ is the dimension of the input vector. In the first ST-block, the input feature sequence is the historical demand series, i.e., $h_{m,i} = x_{m,i}^{T-1}$, $h_{m,i} \in \mathbb{R}^{T \times 2}$. Given $h_{m,i}$, the TCN layer is formulated as:

$$h_{m,i}^{(c)} = (W_{c,1}^{(m)} \ast h_{m,i} + b_{c,1}^{(m)}) \odot \sigma(W_{c,2}^{(m)} \ast h_{m,i} + b_{c,2}^{(m)}), (2)$$

where $\ast$ denotes the convolution operation, $\odot$ is the element-wise multiplication and $\sigma(\ast)$ represents the sigmoid function. The model parameters $W_{c,1}^{(m)}$, $b_{c,1}^{(m)}$, $W_{c,2}^{(m)}$, $b_{c,2}^{(m)}$ are used for feature transformation of nodes in mode $m$ and $W_{c,1}^{(m)}$, $b_{c,1}^{(m)}$, $W_{c,2}^{(m)}$, $b_{c,2}^{(m)}$ are used to compute the output gate for mode $m$. $h_{m,i}^{(c)} \in \mathbb{R}^{K' \times c'}$ is the learned representations for node $i$ in mode $m$ from the TCN layer, where $K'$ is the length of the output feature sequence and $c'$ is the dimension of the output vector.

The learned temporal features from the TCN layers are fed into graph convolution layers to consider spatial and temporal dependencies simultaneously. To encode spatial dependencies among stations/zones within the same mode, we define an intra-modal graph for each mode $m$. Taking bike sharing as an example, its intra-modal graph is defined as $G_b = (V_b, A_b)$, where $V_b$ is a set of BSS stations, and $A_b \in \mathbb{R}^{N_b \times N_b}$ is an adjacency matrix representing the spatial dependencies between BSS stations. Similarly, an intra-modal graph is defined for subway and ride-hailing, denoted as $G_s$ and $G_h$. Prior studies have demonstrated that strong correlations may occur between locations that are either geographically close or semantically similar (i.e., demand patterns) [3]. To encode
both relationships, we define two adjacency matrices for each graph: one for geographical proximity denoted as $A_G$, and the other for pattern similarity denoted as $A_P$. With the geographic center of each zone and station, $A_G$ is computed as:

$$A_{G,ij} = \begin{cases} \exp\left(-\frac{(d_{ij})^2}{\sigma_d}\right) & d_{ij} \leq \kappa_d, \\ 0 & d_{ij} > \kappa_d, \end{cases} \quad (3)$$

where $A_{G,ij}$ is the weight of geographical proximity between nodes $i$ and $j$, $d_{ij}$ is the line distance between $i$ and $j$, $\kappa_d$ is the distance threshold and $\sigma_d$ is the standard deviation of distances. $A_P$ is given as:

$$A_{P,ij} = \frac{\text{Corr}(p_i, p_j)}{\sigma_{p,i}\sigma_{p,j}}, \quad (4)$$

where $A_{P,ij}$ indicates the weight of functional similarity between nodes $i$ and $j$, $p_i$ and $p_j$ are the historical demand series of nodes $i$ and $j$, $\text{Corr}(\cdot)$ calculates the correlation coefficient of two time series vectors, and $\sigma_{p,i}$ and $\sigma_{p,j}$ are the standard deviations of $p_i$ and $p_j$ respectively.

For each mode $m$, given the intra-modal graph $G_m$ and the temporal features learned from the TCN layer $H_m^{(s)} \in \mathbb{R}^{m \times c'}$, the correlations between connected stations within mode $m$ are modeled as:

$$H_m^{(s)} = \text{ReLU}(\tilde{A}_m H_m^{(c)} W_m^{(s)} + l_m^{(s)}), \quad (5)$$

where $W_m^{(s)}$ and $l_m^{(s)}$ are mode-specific model parameters, $\tilde{A}_m = \frac{A_m}{\text{rowsum}(A_m)}$ denotes the normalized adjacency matrix constructed from $A_m$. The spatiotemporal features of mode $m$ are then computed as the sum of temporal and spatial representations:

$$H_m = H_m^{(c)} + H_m^{(s)}, \quad (6)$$

where $H_m$ denotes the spatiotemporal representations of mode $m$ and will be used as input of MRGNN.

2) Domain Discriminator: The domain discriminator is formulated as a multi-class classifier, which uses the learned representation from the spatiotemporal feature extractor as input and maps it to a probability vector denoting which mode the representation is from. Specifically, we first mix the learned representations from different modes. For each node $i$ of mode $m$, the learned feature matrix from the spatiotemporal feature extractor $h_{m,i}$ is then flattened to a 1-d vector, denoted as $\tilde{h}_{m,i}$. Given $\tilde{h}_{m,i}$, we predict the mode it comes from using a feed-forward network:

$$d_{i,1} = \text{ReLU}(W_{d,1} \tilde{h}_{m,i} + b_{d,1}), \quad (7)$$

$$\hat{d}_i = \text{softmax}(W_{d,2}d_{i,1} + b_{d,2}), \quad (8)$$

where $W_{d,1}$, $W_{d,2}$ are the parameter matrices for linear transformation and $b_{d,1}$, $b_{d,2}$ are the biased terms. The output $\hat{d}_i \in \mathbb{R}^3$ is a predicted probability vector denoting which mode the node $i$ is from. The domain discriminator is optimized by minimizing the cross-entropy loss:

$$L_{adg} = - \sum_{m \in \{b, s, h\}} \sum_{i \in m} o_{m,i} \log(\hat{d}_i), \quad (9)$$

where $o_{m,i}$ is a one-hot vector denoting the real mode that the node $i$ is from. Meanwhile, the spatiotemporal feature extractors should be trained to maximize $L_{adg}$ in order to learn transferable temporal features across modes. This is achieved by inserting a gradient reversal layer (GRL) [29] between spatiotemporal feature extractors and the domain discriminator in each STAAN. During backward propagation, GRL reverses the gradient from the domain discriminator and passes it to the preceding GCN and TCN layers. In this way, the spatiotemporal feature extractors and the domain discriminator can be optimized simultaneously with simple backpropagation.

B. Multi-Relational Graph Neural Network

In this subsection, we provide more details on MRGNN that is used to capture the interactions between spatial units across modes. As shown in Figure 2, MRGNN is composed of two major parts: multi-relational graph construction to encode cross-mode spatial dependencies and multi-relational graph convolutions to capture correlations between nodes through message passing.

1) Multi-Relational Graph Construction: As subway and BSS (in our case) are station-based and ride-hailing is stationless, it is difficult to model spatial dependencies on a single homogeneous graph. To encode cross-mode spatial dependencies across modes, we define two inter-modal graphs to capture the pairwise correlations among stations/zones between the target mode (i.e., bike sharing) and each of the auxiliary modes (i.e., subway and ride-hailing). The inter-modal graph between bike sharing and subway is represented as $G_{bs} = (V_b, V_s, A_{bs})$, where $V_b$ and $V_s$ denote BSS and subway stations and $A_{bs} \in \mathbb{R}^{N_b \times N_s}$ is a weighted matrix indicating the cross-mode dependencies between adjacent BSS and subway stations. Similarly, an inter-modal graph is defined between bike sharing and ride-hailing denoted as $G_{bh}$. Similar to inter-modal graphs, we define two adjacency matrices for each inter-modal graph considering both geographical proximity and mobility pattern similarity.

Figure 3 illustrates the constructed intra- and inter-modal graphs among bike sharing, subway and ride-hailing. A total of $(3+2) \times 2 - 10$ relations is defined to encode spatial dependencies between nodes from different modes, including 3 intra-modal and 2 inter-modal graphs, each with 2 adjacency matrices. This formulation can be easily adapted to encode spatial dependencies across other modes with diverse network structures.

2) Multi-Relational Graph Convolutions: Graph convolutions have been an effective way to aggregate information about connected nodes. However, most existing GCNs cannot be applied to the multi-relational graph defined above since...
they cannot process inter-modal graphs with different types of nodes and a non-square adjacency matrix. To tackle these issues, we adopt an inter-modal graph convolution network introduced in [13]. It takes the learned representations of stations and zones from the STAAN module as input and aggregates information of connected subway stations and ride-hailing zones for each BSS station considering the effect of spatial dependencies. Mathematically, for each time step, given the inter-modal graphs \( G_{bs} = (V_b, V_s, A_{bs}) \) and \( G_{bh} = (V_b, V_h, A_{bh}) \), as well as the representations of subway and ride-hailing learned from STAAN, denoted as \( H_b \in \mathbb{R}^{N_b \times c} \), \( H_s \in \mathbb{R}^{N_s \times c'} \) respectively, the inter-modal relationship is modeled as:

\[
Z_{bs} = ReLU(\tilde{A}_b(\tilde{A}_s H_s) W_{bs} + l_{bs}), \quad (10)
\]

\[
Z_{bh} = ReLU(\tilde{A}_b(\tilde{A}_h H_h) W_{bh} + l_{bh}), \quad (11)
\]

where \( Z_{bs} \in \mathbb{R}^{N_b \times c'} \), \( Z_{bh} \in \mathbb{R}^{N_b \times c''} \) are the aggregated features of connected subway stations and ride-hailing zones on BSS stations respectively, \( c'' \) represents the output vector dimension of each node (i.e. station/zone) from the GCN layer. \( W_{bs}, W_{bh} \) and \( l_{bs}, l_{bh} \) are the learned model parameters. \( \tilde{A} = \frac{A}{\text{rowsum}(A)} \) denotes the normalized adjacency matrix constructed from \( A \).

In addition, we apply an intra-modal GCN layer to model pairwise correlations between connected BSS stations. In practice, the inter- and intra-modal graph convolution layers can be modeled in parallel using batch matrix multiplication operations. Through the intra- and inter-modal graph convolutions, each BSS station receives multiple feature vectors from geographically adjacent or semantically similar subway stations, ride-hailing zones and other BSS stations. The learned feature vectors from heterogeneous neighborhood nodes are then aggregated using an adding function.

### C. Prediction Layer

Based on the aggregated multi-modal spatiotemporal features from MRGNN, we generate bike sharing demand predictions at the next time step using a two-layer fully-connected feed-forward network. The prediction error is given by:

\[
L_{pre} = \sum_{i \in V_b} ||\hat{x}_{b,i}^{t+1} - x_{b,i}^{t+1}||^2, \quad (12)
\]

where \( x_{b,i}^{t+1} \) and \( \hat{x}_{b,i}^{t+1} \) are the predicted and true demand values of BSS station \( b \) at time step \( t + 1 \). The final loss function contains two parts: prediction loss of BSS demand \( L_{pre} \) and adversarial loss of the domain discriminator \( L_{adv} \). The overall loss function is defined as follows:

\[
L = L_{pre} + \epsilon_{adv} L_{adv}, \quad (13)
\]

where \( \epsilon_{adv} \) is used to trade off different parts of loss.

### V. Experimental Setup

#### A. Data Description

Our proposed model is validated on real-world public datasets from NYC. In this study, we use Manhattan as the research area and collect travel demand data of bike sharing, subway and ride-hailing from 2018-03-01 to 2018-08-31. Specifically, we use the following three datasets:

- **NYC Citi Bike**: the data consists of the pick-up and drop-off time and station of each Citi Bike trip records. During our study period, there are around 36 thousand trip records per day. The original dataset contains 436 BSS stations. We filter out the stations with an average of fewer than three orders per hour and keep 246 BSS stations for demand prediction. In Appendix A, we also show our model results for all 436 stations.

- **NYC Subway**: the data contains the entries and exits counts of each turnstile in subway stations every four hours, with around 2.4 million entry/exit counts every day on average. We filter out subway stations with no demand for long time periods, which results in 107 subway stations.

- **NYC Ride-hailing**: we use the for-hire vehicle (FHV) data from NYC Taxi & Limousine Commission (TLC), which is provided by ride-hailing companies such as Uber and Lyft. It contains the pick-up and drop-off time and zone of each trip records during the study period. On average there are 234 thousand trips per day during our study period. The zones are pre-determined by TLC and there are 63 TLC zones in Manhattan.

We present the spatial distribution and temporal pattern of the multimodal travel demand in NYC in Fig. 4 and Fig. 5. For all modes, stations/zones with intense demand are concentrated in Midtown Manhattan, followed by Downtown. This indicates that the demand of different modes exhibits similar spatial distributions, supporting our idea of using inter-modal demand to enhance the prediction performance of bike sharing. From Fig. 5, we find that there is a notable disparity between the demand patterns of different modes. The usage of bike sharing is more active during 08:00-20:00 especially at evening peak (16:00-20:00), while ride-hailing is busier at night (16:00-24:00). On the other hand, the outflow and inflow demand of subway exhibits different patterns: its inflow has two peaks in the morning and evening, while its outflow is concentrated in the evening peak. This is likely because most people take the subway to Manhattan for work in the morning and leave Manhattan for home in the evening. Such results support the necessity to mitigate the effect of different temporal patterns when leveraging cross-mode information.
Fig. 4. Spatial distribution of multimodal travel demand in Manhattan.

Fig. 5. Temporal pattern of multimodal travel demand in Manhattan. For a more intuitive comparison, we show the relative proportion of travel demand for different time periods to the total daily travel demand.

B. Experiment Settings

We compare DA-MRGNN with two groups of baselines. The first group contains methods with its own historical demand as input:

- **XGBoost**: a machine learning model capturing the relationship between future demand and historical demand series based on gradient boosting trees;
- **LSTM**: a variant of RNNs using memory units to capture long-term temporal dependencies;
- **STGCN** [4]: a framework using GCNs for spatial dependencies and TCNs for temporal dependencies;
- **MGCN** [3]: a GCN-based model using multiple graphs to encode spatial correlations and conceptual RNNs to model temporal correlations;
- **GWNET** [23]: a graph learning approach which captures the hidden spatial dependencies in the demand data by learning an adjacency matrix through node embedding.

The second group contains BSS demand prediction methods considering the effect of other modes:

- **MM-STGCN** [9]: a STGCN-based approach which incorporates the usage of auxiliary modes as additional features of BSS stations for model input. For each BSS station, the usage features of subway and ride-hailing zones are computed as a weighted sum of the usage record of subway stations and ride-hailing zones, weighted by distance to the BSS station.
- **MM-GWNET**: Following the idea of MM-STGCN, we implement a GWNET-based model with the accumulated demand of subway stations and ride-hailing zones as additional input.
- **ST-MRGNN** [13]: a demand co-prediction method for multimodal systems with multi-relational graph neural networks to capture the spatial dependencies across different modes. In our implementation, this method is adapted to jointly predict the demand of bike sharing, subway and ride-hailing.

For fair comparison, we use the same experiment settings for all models. Specifically, the demands of all three modes are aligned into 4-hour intervals because the NYC subway demand data has a temporal resolution of 4 hours. To explore the potential effect of higher temporal resolution, we also conduct experiments using a prediction interval of 1 hour with NYC ride-hailing and bike sharing data as inputs, and the results are shown in Appendix B. Min-max normalization is used for each mode to mitigate the effect of demand variance. We set the historical time step $T = 6$ and predict demand at the next time step. Data from the first 60% time steps are used for model training, the following 20% for validation, and the last 20% for model testing. For all DNN-based models, we use 500 training epochs, with the learning rate of 0.002, the batch size of 32 and the dropout ratio of 0.3. To prevent overfitting, we use early stopping of 100 epochs on the validation set and a L2 regularization on the loss function with a weight decay of 1e-5. All models are evaluated using three widely used metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Coefficient of Determination ($R^2$). We run 5 independent experiments for each model and report the average values on the test set.

VI. Results

A. Comparison of Model Performance

In this subsection, we compare the prediction performance of our proposed model with baselines on NYC bike sharing data. The performance of different models are summarized in Table I. Compared with the baseline models, our proposed model achieves significantly superior performance for all evaluation metrics. This is likely because our model can directly leverage spatiotemporal information of related subway stations and ride-hailing zones to enhance the prediction of bike sharing demand. In addition, our model reduces the effect of distribution discrepancy between modes with adversarial learning. Figure 6 presents the performance variance of 5 independent runs of our proposed model and selected DNN-based baselines. We can find that our proposed model has the best performance regarding both RMSE and MAE in all experiments with relatively high stability. In Appendix C, we further analyze the performance of our model for different time periods and stations, as well as under anomaly conditions.

Among baseline models, the poor performance of LSTM can be potentially explained by its inability to capture spatiotemporal dependencies in the demand data. Figure 6 presents the performance variance of 5 independent runs of our proposed model and selected DNN-based baselines. We can find that our proposed model has the best performance regarding both RMSE and MAE in all experiments with relatively high stability. In Appendix C, we further analyze the performance of our model for different time periods and stations, as well as under anomaly conditions.

Among baseline models, the poor performance of LSTM can be potentially explained by its inability to capture spatiotemporal dependencies in the demand data.
TABLE I  

| Input   | Models          | RMS E | MAE  | $R^2$ |
|---------|-----------------|-------|------|-------|
| Single-mode | XGBoost        | 15.139 | 9.262 | 0.734 |
|          | LSTM           | 16.148 | 10.255 | 0.695 |
|          | STGCN          | 13.829 | 9.097 | 0.789 |
|          | MGSTGCN        | 14.436 | 9.359 | 0.760 |
|          | GWNET          | 12.689 | 8.092 | 0.814 |
|          | GMSDR          | 13.536 | 9.065 | 0.790 |
| Multi-mode | MM-STGCN      | 14.078 | 8.953 | 0.780 |
|          | MM-GWNET       | 12.461 | 7.998 | 0.820 |
|          | ST-MRGNN       | 11.849 | 7.329 | 0.843 |
|          | DA-MRGNN       | 11.170 | 7.053 | 0.859 |

**Fig. 6.** Comparison of model stability.

inter-modal dependencies between BSS stations. XGBoost performs notably better than LSTM, showing the advantage of gradient boosting machines in some cases. Although MGSTGCN encodes multiple types of spatial dependencies, it performs worse than STGCN in our experiments, which might be due to its low model stability as illustrated in Figure 6. GMSDR performs better than STGCN, which can be explained by the novel recurrent structure to capture temporal relationships. Meanwhile, GWNET can further improve the prediction performance compared to the other single-mode baselines given its ability to learn spatial dependencies hidden in demand data with an adaptive adjacency matrix. Compared with STGCN, MM-STGCN performs even worse regarding RMSE and $R^2$ in our case. MM-GWNET performs better than GWNET, though the improvement is marginal. This suggests that incorporating multi-modal knowledge as additional attributes of BSS stations might not be able to make significant contributions to the model performance and may even lead to negative transfer. ST-MRGNN can notably outperform the other baseline models due to its ability to leverage spatiotemporal relationships from heterogeneous spatial units of different modes. Compared with ST-MRGNN, our proposed model can further reduce the prediction error with an RMSE improvement of 5.7%.

B. The Effect of Input Modes

To further investigate the effect of incorporating inter-modal relationships on bike sharing demand prediction, we compare the performance of our model using different input mode combinations: bike sharing alone, bike sharing and subway, bike sharing and ride-hailing, and all three modes. The results are shown in Table II. Note that they share the same model structure, just with different input modes. We find that the

**TABLE II**  

| Models                  | RMS E  | MAE  | $R^2$  |
|-------------------------|--------|------|--------|
| Bike sharing            | 11.781 | 7.470 | 0.841  |
| Bike sharing + subway   | 11.508 | 7.240 | 0.848  |
| Bike sharing + ride-hailing | 11.315 | 7.123 | 0.853  |
| Bike sharing + subway + ride-hailing | 11.170 | 7.053 | 0.859  |

**Fig. 7.** Comparison of DA-MRGNN and S-MRGNN (a variant of DA-MRGNN with single-mode input) at different time and stations.

inter-modal relations from geographically nearby or semantically similar subway stations and ride-hailing zones can indeed help with the prediction of bike sharing demand: without any inter-modal relations, the prediction error regarding RMSE and MSE would increase by 5.2% and 5.6% respectively. Incorporating either subway or ride-hailing demand patterns can already significantly improve the prediction performance of bike sharing demand. It is also worth noting that with only bike sharing demand as input, our proposed model can still perform better than the single-mode version of GWNET, suggesting that our spatiotemporal framework can better extract intramodal relationships.

To investigate how the effect of inter-modal relationships varies over space and time, we compare our proposed model with a variant of DA-MRGNN using single-mode demand as input (S-MRGNN) at different time and stations. The results are displayed in Figure 7. It can be clearly seen from Figure 7 that DA-MRGNN performs better than S-MRGNN at all time intervals. The advantage of DA-MRGNN over S-MRGNN is especially significant during 16:00-20:00, indicating that our model can successfully capture the complex relations between different modes during rush hours. During 04:00-08:00, DA-MRGNN also performs notably better than S-MRGNN. This suggests that our model can effectively utilize the knowledge from other modes to improve the prediction accuracy of demand-sparse time periods. Figure 7b shows that the prediction performance of almost all stations can benefit from the cross-mode knowledge, especially for several BSS stations in Midtown and Downtown Manhattan. In summary, inter-modal relations can contribute to the demand prediction performance of bike sharing across different time and space.

C. The Effect of Adversarial Learning

In this subsection, we conduct ablation experiments to quantify the contribution of adversarial learning in our proposed model. Specifically, we implement a model variant with the
domain discriminator module dropped and the model is trained with only $L_{\text{pre}}$. The averaged results over 5 runs are shown in Table III. We can see that removing the adversarial learning module would lead to a notable increase in the prediction error. To assess the confidence interval of such improvement, we use a $t$-test to confirm the statistical significance of such performance difference (see Appendix D for more details). We further compare the computation costs in Table IV, and the results show that adversarial learning does not require much additional computation time for both model training and inference. To investigate the stability of the model performance, we can also compare the effect of adversarial learning on model performance in various experiment settings. In Appendix A, the comparative results of DA-MRGNN with and without adversarial learning are shown for all 436 BSS stations in the NYC Citi Bike data. In Appendix B, the two models are compared with a prediction interval of 1 hour. All the results validate the value of adversarial learning for improved model performance. Therefore, we can reasonably conclude that adversarial learning consistently and significantly contributes to BSS demand prediction performance with little additional computation cost.

To further investigate how the trade-off between $L_{\text{pre}}$ and $L_{\text{adv}}$ influences the model performance, we change $\epsilon_{\text{adv}}$ from 0 to 2 with a step size of 0.5. The results are displayed in Figure 8. It is found that the model performance is relatively good with $\epsilon_{\text{adv}}$ ranging from 0.5 to 1.5, whereas the performance degrades as $\epsilon_{\text{adv}}$ gets larger. This indicates that, rather than completely eliminating the distribution discrepancy between modes, it can be helpful to preserve some meaningful inter-modal differences. Also, a large $\epsilon_{\text{adv}}$ may lead to the model over-emphasizing on $L_{\text{adv}}$, which can be detrimental to the overall prediction performance.

### Table III

| Models     | RMSE      | MAE     | $R^2$  |
|------------|-----------|---------|--------|
| - adversarial | 11.454    | 7.209   | 0.850  |
| DA-MRGNN   | 11.170    | 7.053   | 0.859  |

### Table IV

| Models     | Training Time (s/epoch) | Inference Time (s) |
|------------|-------------------------|--------------------|
| - adversarial | 0.880                   | 0.120              |
| DA-MRGNN   | 0.917                   | 0.124              |

D. Interpretation Analysis

In this section, we adapt an explainable GNN technique namely GNNEExplainer [33] to understand how DA-MRGNN utilizes intra- and inter-modal relationships to make predictions. Briefly, it provides local explanation for the prediction of each BSS station by identifying a subset of connected subway stations, ride-hailing zones and BSS stations that are important to its prediction. More details of the GNNEExplainer technique used in our study are introduced in Appendix E. Fig. 9 shows the interpretation results for a few example BSS stations. It can be seen that the target BSS station in Fig. 9(a) is mainly affected by an adjacent subway station. Similarly, the target BSS station in Fig. 9(b) is mainly determined by the ride-hailing zone it falls in. These two examples verify the effectiveness of considering spatially adjacent subway stations and ride-hailing zones for bike sharing demand prediction. In addition, a BSS station can also be highly affected by stations or zones that are not in its neighborhood, as shown in Fig. 9(c). Based on this example, we further analyze how distant stations and zones can affect a target BSS station. It is found that among all BSS stations, Pershing Square North has the strongest OD connection with the target station. The Time Sq.–42nd St. subway station has high pattern similarity with the target BSS station with a correlation of 0.869. This indicates that distant stations/zones can also contribute to demand prediction due to similar pattern distributions or strong OD connections.

### VII. Conclusion

Bike sharing demand prediction is quite crucial for the efficient operation of bike sharing systems. In this paper, we aim to enhance the prediction performance of bike sharing by incorporating cross-mode information. This problem is challenging regarding how to fuse multimodal spatiotemporal information and how to handle the distribution discrepancy of demand patterns between modes. To address these issues, a domain-adversarial graph neural network is proposed to extract shareable features from spatial units of different modes. To extract shareable spatiotemporal features, a spatiotemporal adversarial adaptation network is developed by taking advantage of adversarial learning. The spatial dependencies
across modes are encoded with multiple intra- and inter-modal graphs, and an inter-modal graph convolution layer is introduced to capture correlations between nodes from different modes. Empirical validation is performed on real-world bike sharing, subway and ride-hailing datasets from NYC. The results show that our proposed model achieves the best performance compared to existing methods, suggesting that the knowledge of subway and ride-hailing demand can indeed benefit the demand prediction of bike sharing. Further analysis demonstrates the effectiveness of all the proposed components.

In future works, this research can be improved in the following aspects. Although our study predominantly focuses on station-based bike sharing due to data constraints, our model boasts adaptability. By adjusting the spatial unit definition and updating the adjacency matrices, our model can be repurposed for dockless BSS systems and other modes of transportation. For this study, we have chosen an adaptable adjacency matrix definition that is compatible with all spatial units. Looking forward, there is potential for future research to investigate bespoke adjacency matrix definitions tailored for specific spatial units, thus optimizing predictions across diverse transportation systems. Secondly, our current research is based on a half-year dataset and does not explicitly consider the dynamic variations in inter-modal relationships across different seasons. For instance, during the summer, individuals may use bike-sharing to complement their public transit usage, while public transit can be competitive with bike-sharing in the winter as people opt for warmer travel alternatives. Upcoming studies could delve into how these inter-modal relationships evolve monthly and how to better capture such seasonal variations, thereby enhancing cross-mode knowledge adaptation. Thirdly, it is demonstrated in Appendix C that our model performance can be limited under anomaly conditions. To address this issue, it would be meaningful to develop demand prediction models that can work well even for rare events (e.g., big festivals, road closure, etc.). This problem is challenging due to different data distributions under anomaly conditions and data sparsity for model training. A potential solution is to introduce statistical models in spatiotemporal deep learning frameworks, as recently proposed in [34].

**APPENDIX A**

**Prediction Results for All BSS Stations**

In this section, we show the comparative results of different models for all BSS stations in NYC Citi Bike data. Table V shows the averaged performance of DA-MRGNN and two selected baselines over 5 independent runs. Specifically, we use GWNET as a representative of single-mode baselines and ST-MRGNN as a representative of multi-mode baselines. It can be seen that ST-MRGNN performs better than GWNET as it leverages multi-modal information.

| Model | RMSE | MAE | $R^2$ |
|-------|------|-----|-------|
| Single-mode | GWNET | 10.533 | 6.582 | 0.829 |
| Multi-mode | ST-MRGNN | 9.838 | 5.801 | 0.856 |
| Multi-mode | DA-MRGNN | 9.227 | 5.391 | 0.868 |

**APPENDIX B**

**Prediction Results With 1-Hour Time Intervals**

In this section, we test our model performance with 1-hour time intervals using ride-hailing and bike sharing data as input. Specifically, we use the historical demand for ride-hailing and bike sharing in the past 6 hours to predict BSS demand in the next hour. We compare DA-MRGNN with GWNET and ST-MRGNN in Table VII and two model variants without either cross-mode information (cross-mode) and the other without adversarial learning (adversarial). It can be found that both cross-mode information and adversarial learning contribute to the improved model performance.

**APPENDIX C**

**Spatiotemporal Analysis of DA-MRGNN**

In this section, we analyze how our model performance varies for different stations and time as well as under anomaly conditions. Firstly, we compare the performance of our model for high-demand and low-demand stations and the results are displayed in Table IX. Specifically, the top 1/3 of BSS stations with the highest average demand are labeled as high-demand stations and the bottom 1/3 as low-demand stations. It is found that the high-demand stations have a higher $R^2$ on average than low-demand stations, which is potentially because high-demand stations are more predictable with less uncertainty. Secondly, we compare the performance of DA-MRGNN during peak (i.e. 08:00-20:00) and non-peak

| Model | RMSE | MAE | $R^2$ |
|-------|------|-----|-------|
| Single-mode | GWNET | 4.778 | 2.950 | 0.704 |
| Multi-mode | ST-MRGNN | 4.221 | 2.566 | 0.768 |
| Multi-mode | DA-MRGNN | 4.145 | 2.529 | 0.777 |

| Model | RMSE | MAE | $R^2$ |
|-------|------|-----|-------|
| Single-mode | GWNET | 4.265 | 2.631 | 0.766 |
| Multi-mode | ST-MRGNN | 4.164 | 2.538 | 0.774 |
| Multi-mode | DA-MRGNN | 4.145 | 2.529 | 0.777 |

| Model | RMSE | MAE | $R^2$ |
|-------|------|-----|-------|
| Single-mode | GWNET | 5.627 | 5.655 | 0.835 |
| Multi-mode | ST-MRGNN | 9.481 | 5.551 | 0.859 |
| Multi-mode | DA-MRGNN | 9.227 | 5.391 | 0.868 |

| Input | Models | RMSE | MAE | $R^2$ |
|-------|-------|------|-----|-------|
| Single-mode | GWNET | 4.778 | 2.950 | 0.704 |
| Multi-mode | ST-MRGNN | 4.221 | 2.566 | 0.768 |
| Multi-mode | DA-MRGNN | 4.145 | 2.529 | 0.777 |

| Input | Models | RMSE | MAE | $R^2$ |
|-------|-------|------|-----|-------|
| Single-mode | GWNET | 10.533 | 6.582 | 0.829 |
| Multi-mode | ST-MRGNN | 9.838 | 5.801 | 0.856 |
| Multi-mode | DA-MRGNN | 9.227 | 5.391 | 0.868 |
TABLE IX
PERFORMANCE COMPARISON BETWEEN HIGH-DEMAND AND LOW-DEMAND STATIONS

| Stations          | RMSE | MAE  | $R^2$  |
|-------------------|------|------|--------|
| High-demand       | 14.967 | 9.223 | 0.856  |
| Low-demand        | 7.681  | 5.425 | 0.788  |

TABLE X
PERFORMANCE COMPARISON BETWEEN PEAK AND NON-PEAK HOURS

| Time   | RMSE | MAE  | $R^2$  |
|--------|------|------|--------|
| Peak   | 14.465 | 9.923 | 0.799  |
| Non-peak | 6.180  | 4.150 | 0.775  |

![Graph showing model performance under anomaly conditions.]

Fig. 10. Model performance under anomaly conditions.

APPENDIX D
THE $t$-TEST RESULT OF ADVERSARIAL LEARNING

We use $t$-test to assess whether the contribution of adversarial learning is statistically significant. The results are shown in Table XI. It can be found that in different experiment settings, adversarial learning can consistently improve the performance of all evaluation metrics with more than 90% confidence level.

TABLE XI
$t$-TEST RESULT OF PERFORMANCE DIFFERENCE WITH AND WITHOUT ADVERSARIAL LEARNING

| experiments | $t$-statistic | p-value | $R^2$   |
|-------------|---------------|---------|---------|
| Section VI-C| -1.648*       | 0.087   | 2.241** |
| Appendix A  | -2.926***     | 0.008   | 3.288***|
| Appendix B  | -3.151**      | 0.017   | 3.672** |

** and *** represent the confidence level at 90%, 95%, 99% respectively.

APPENDIX E
GNNEPARATOR FOR MODEL INTERPRETATION

GNNEPARATOR [33] is one of the pioneering works for explainable GNNs, whose main idea is to identify a subgraph that maximizes its mutual information with the GNN’s prediction. The underlying assumption is that if the prediction result of the target node is largely determined by a connected node, their dependencies should be strong. The subgraph is identified by formulating a mean field variational approximation and learning a real-valued graph mask to select important edges. In our case, we construct a node mask for each mode in the multimodal network so that our proposed framework can identify important intra- and inter-modal relationships simultaneously. Mathematically, for a target BSS station $i$, we randomly initialize a node mask for bike sharing, subway and ride-hailing, respectively denoted as $M_{b,i} \in \mathbb{R}^{E_b}$, $M_{s,i} \in \mathbb{R}^{E_s}$, $M_{h,i} \in \mathbb{R}^{E_h}$. The masks are optimized using:

$$
\min_{\{M_{b,i}, M_{s,i}, M_{h,i}\}} \text{dist}(F_i(x_{b,i}^{t-T:T}, x_{s,i}^{t-T:T}, x_{h,i}^{t-T:T})),
$$

$$
F_i(x_{b,i}^{t-T:T}, x_{s,i}^{t-T:T}, x_{h,i}^{t-T:T})) = \sigma(M_{b,i}),
$$

$$
X_{s,i}^{t-T:T} = X_{s,i}^{t-T:T} \circ \sigma(M_{s,i}),
$$

$$
X_{h,i}^{t-T:T} = X_{h,i}^{t-T:T} \circ \sigma(M_{h,i}).
$$

We use L2 norm as the distance function $\text{dist}(\cdot)$. The learned node masks can effectively select a small subset of subway stations, ride-hailing zones and BSS stations that are important to the GNN’s prediction of a target BSS station.

REFERENCES

[1] C. Xu, J. Ji, and P. Liu, “The station-free sharing bike demand forecasting with a deep learning approach and large-scale datasets,” Transp. Res. C, Emerg. Technol., vol. 95, pp. 47–60, Oct. 2018.
[2] D. Liu, J. Wang, S. Shang, and P. Han, “MSDR: Multi-step dependency relation networks for spatial temporal forecasting,” in Proc. 28th ACM SIGKDD Conf. Knowl. Discovery Data Mining, Aug. 2022, pp. 1042–1050.
[3] G. Xu et al., “Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting,” in Proc. AAAI Conf. Artif. Intell., Worcester, MA, USA, 2019, vol. 33, no. 1, pp. 3656–3663.
[4] B. Yu, H. Yin, and Z. Zhu, “Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting,” 2017, arXiv:1709.04875.
[5] C. Zhang, L. Zhang, Y. Liu, and X. Yang, “Short-term prediction of bike-sharing usage considering public transport: A LSTM approach,” in Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC), Nov. 2018, pp. 1564–1571.
[6] J. Ye, L. Sun, B. Du, Y. Fu, X. Tong, and H. Xiong, “Co-prediction of multiple transportation demands based on deep spatio-temporal neural network,” in Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Jul. 2019, pp. 305–313.
[7] S. Wang, H. Miao, H. Chen, and Z. Huang, “Multi-task adversarial spatial–temporal networks for crowd flow prediction,” in Proc. 29th ACM Int. Conf. Inf. Knowl. Manage., Oct. 2020, pp. 1555–1564.
[8] H. Xu, T. Zou, M. Liu, Y. Qiao, J. Wang, and X. Li, “Adaptive spatiotemporal dependence learning for multi-mode transportation demand prediction,” IEEE Trans. Intell. Transp. Syst., vol. 23, no. 10, pp. 18632–18642, Oct. 2022.
[9] J.-H. Cho, S. W. Ham, and D.-K. Kim, “Enhancing the accuracy of peak hourly demand in bike-sharing systems using a graph convolutional network with public transit usage data,” Transp. Res. Rec., J. Transp. Res. Board, vol. 2675, no. 10, pp. 554–565, Oct. 2021.
[10] C. Li, L. Bai, W. Liu, L. Yao, and S. T. Waller, “A multi-task memory network with knowledge adaptation for multimodal demand forecasting,” Transp. Res. C, Emerg. Technol., vol. 131, Oct. 2021, Art. no. 103352.
[11] J. W. Yoon, F. Pinelli, and F. Calabrese, “Cityride: A predictive bike-sharing trip model,” Transp. Res. C, Emerg. Technol., vol. 133, Dec. 2021, Art. no. 103404.

[12] M. Hua, F. Camara Pereira, Y. Jiang, and X. Chen, “Transfer learning for cross-modal demand prediction of bike-share and public transit,” 2022, arXiv:2203.09279.

[13] Y. Liang, G. Huang, and Z. Zhao, “Joint demand prediction for multimodal systems: A multi-task multi-relational spatiotemporal graph neural network approach,” Transp. Res. C, Emerg. Technol., vol. 140, Jul. 2022, Art. no. 103731.

[14] Y. Liang, G. Huang, and Z. Zhao, “Bike sharing demand prediction based on knowledge sharing across modes: A graph-based deep learning approach,” in Proc. IEEE 25th Int. Conf. Intell. Transp. Syst. (ITSC), Oct. 2022, pp. 857–862.

[15] J. W. Yoon, F. Pinelli, and E. Calabrese, “Cityride: A predictive bike sharing journey advisor,” in Proc. IEEE 13th Int. Conf. Mobile Data Manage., Jul. 2012, pp. 306–311.

[16] C. Rudloff and B. Lackner, “Modeling demand for bikesharing systems: Neighboring stations as source for demand and reason for structural breaks,” Transp. Res. Rec., J. Transp. Res. Board, vol. 2430, no. 1, pp. 1–11, Jan. 2014.

[17] S. He, H. Chen, C. Du, J. Li, and N. Jing, “A hierarchical demand prediction method with station clustering for bike sharing system,” in Proc. IEEE 3rd Int. Conf. Data Sci. Cyberspace (DSC), Jun. 2018, pp. 829–836.

[18] S. Guidon, D. J. Reck, and K. Axhausen, “Expanding a(n) (electric) bicycle-sharing system to a new city: Prediction of demand with spatial regression and random forests,” J. Transp. Geography, vol. 84, Apr. 2020, Art. no. 102692.

[19] J. Ke et al., “Hexagon-based convolutional neural network for supply-demand forecasting of ride-sourcing services,” IEEE Trans. Intell. Transp. Syst., vol. 20, no. 11, pp. 4160–4173, Nov. 2019.

[20] Y. Liang, F. Ding, G. Huang, and Z. Zhao, “Deep trip generation with graph neural networks for bike sharing system expansion,” Transp. Res. C, Emerg. Technol., vol. 154, Sep. 2023, Art. no. 104241.

[21] L. Lin, Z. He, and S. Peeta, “Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach,” Transp. Res. C, Emerg. Technol., vol. 97, pp. 258–276, Dec. 2018.

[22] S. He and K. G. Shin, “Towards fine-grained flow forecasting: A graph attention approach for bike sharing systems,” in Proc. Web Conf., Apr. 2020, pp. 88–98.

[23] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, “Graph WaveNet for deep spatial–temporal graph modeling,” 2019, arXiv:1906.00121.

[24] Y. Wang, Z. Zhan, Y. Mi, A. Sobhani, and H. Zhou, “Nonlinear effects of factors on dockless bike-sharing usage considering grid-based spatiotemporal heterogeneity,” Transp. Res. D, Transp. Environ., vol. 104, Mar. 2022, Art. no. 103194.

[25] A. Farahani, S. Vogheli, K. Rasheedi, and H. R. Arabnia, “A brief review of domain adaptation,” in Advances in Data Science and Information Engineering (Transactions on Computational Science and Computational Intelligence), R. Stabholz, G. M. Weiss, M. Abou-Nassar, C. Y. Yang, H. R. Arabnia, and L. Deligiannidis, Eds. Cham, Switzerland: Springer, 2021, doi: 10.1007/978-3-030-71704-9_65.

[26] Y. Liu et al., “Knowledge transfer with weighted adversarial network for cold-start store site recommendation,” ACM Trans. Knowl. Discovery Data, vol. 15, no. 3, pp. 1–27, Jun. 2021.

[27] E. Teng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, “Deep domain confusion: Maximizing for domain invariance,” 2014, arXiv:1412.3474.

[28] Y. Tang, A. Qu, A. H. F. Chow, W. H. K. Lam, S. C. Wong, and W. Ma, “Domain adversarial spatial–temporal network: A transferable framework for short-term traffic forecasting across cities,” 2022, arXiv:2202.03630.

[29] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 1180–1189.

[30] Z. Fang, D. Wu, L. C. Lu Pan, and Y. Gao, “When transfer learning meets cross-city urban flow prediction: Spatio-temporal adaptation matters,” in Proc. IJCAI, 2022, pp. 2030–2036.

[31] I. Goodfellow et al., “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., vol. 27, 2014, pp. 1–12.

[32] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, “GNNExplainer: Generating explanations for graph neural networks,” in Proc. Adv. Neural Inf. Process. Syst., vol. 32, 2019, pp. 1–10.

[33] Q. Wang, S. Wang, D. Zhuang, H. Koutsopoulos, and J. Zhao, “Uncertainty quantification of spatiotemporal travel demand with probabilistic graph neural networks,” 2023, arXiv:2303.04040.