

Cross-Modality and Equity-Aware Graph Pooling Fusion: A Bike Mobility Prediction Study

Xi Yang* Suining He*† Kang G. Shin Mahan Tabatabaie Jing (David) Dai

Abstract—We propose an equity-aware *GRA*ph-fusion differentiable *P*ooling neural network to accurately predict the spatio-temporal urban mobility (e.g., station-level bike usage in terms of departures and arrivals) with *E*quity (*GRAPE*). *GRAPE* consists of two independent hierarchical graph neural networks for two mobility systems — one as a target graph (i.e., a bike sharing system) and the other as an auxiliary graph (e.g., a taxi system). We have designed a convolutional fusion mechanism to jointly fuse the target and auxiliary graph embeddings and extract the shared spatial and temporal mobility patterns within the embeddings to enhance prediction accuracy. To further improve the equity of bike sharing systems for diverse communities, we focus on the bike resource allocation and model prediction performance, and propose to regularize the predicted bike resource as well as the accuracy across advantaged and disadvantaged communities, and thus mitigate the potential unfairness in the predicted bike sharing usage. Our evaluation of over 23 million bike rides and 100 million taxi trips in New York City and Chicago has demonstrated *GRAPE* to outperform all of the baseline approaches in terms of prediction accuracy (by 15.80% for NYC and 50.55% for Chicago on average) and social equity awareness (by 32.44% and 24.43% in terms of resource fairness for NYC and Chicago, and 13.36% and 16.52% in terms of performance fairness).

Index Terms—Bike sharing, equity awareness, usage prediction, target and auxiliary graphs, differentiable pooling.

1 INTRODUCTION

The operation efficiency and profitability of urban mobility systems (e.g., bike and ride sharing, and other public transportation systems) depends heavily on the mobility modeling for accurate prediction of mobility resource usage, which is key to distributing mobility resources (e.g., rebalancing bike demands and supplies), meeting the mobility needs, and enhancing communities' access to life-essential resources (e.g., grocery, employment opportunities) in the city.

In this paper, we address the following two important questions with the focus on station-based bike sharing as a representative case study to re-imagine the predictive usage modeling and accessibility implications regarding this popular first- and last-mile connectivity (say, between the riders' home locations and the bus stops, or between the metro stations and their employment locations).

(a) How to capture and exploit cross-modality mobility patterns? Existing bike sharing modeling techniques largely consider single modality (i.e., the bike sharing system itself), and focus on engineering and integrating of various exogenous features [1]–[5] for accurate prediction. Different mobility systems or modalities, such as taxi, bus,

and subway, in the transportation network in a city might share similar mobility patterns due to the people's commute and travel preferences. Despite these modalities' seemingly heterogeneous data representations [6]–[8], accounting independently for their mobility patterns will likely discard the essentially useful mobility patterns shared across these platforms, leading to degraded prediction accuracy. Furthermore, existing studies [9], [10] largely focus on hand-crafted feature engineering to integrate the patterns. Such approaches cannot fully extract the complex latent correlations across these modalities. In fact, these mobility systems may be represented or described in terms of network graphs, with their mobility trips, such as bike and taxi rides, connecting different city regions – for instance, bike stations and taxi pick-up/drop-off zones. Such a representation provides excellent feature interpretation regarding the spatial connectivity, and the latent correlations across these graphs of different modalities, may bring complementary knowledge beyond the engineering of single-modality features. However, how to design an effective way to fuse these graphs for the model predictability enhancement remains largely under-explored.

(b) How to account for the equity-aware prediction modeling? We focus on bike sharing as it has great potential for serving the broader communities due mainly to its ease of use and low cost, particularly providing disadvantaged communities accessibility at low cost. However, the recent publications and various field studies [11], [12] have revealed that regions with more economic activities (e.g., central business area, tourism zones) exhibit more bike usage (i.e., departures and arrivals). These regions may often be predominated by historically advantaged communities (e.g., in terms of ethnicity, education, and income levels) who may therefore be allocated with more bike resources and infrastructures, if based on the historical fulfilled mobility

- Xi Yang*, Suining He* (*equal contributions), and Mahan Tabatabaie, are with the School of Computing, University of Connecticut, Storrs, CT. E-mail: {xi.yang, suining.he, mahan.tabatabaie}@uconn.edu
- Kang G. Shin is with the Department of Electrical Engineering and Computer Science, University of Michigan Ann Arbor, MI. E-mail: kgshin@umich.edu
- Jing (David) Dai is with the Google Maps Department, Mountain View, CA. E-mail: jddai@google.com
- Contact Author (†): Suining He (suining.he@uconn.edu)

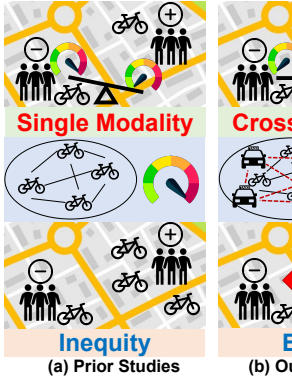


Fig. 1: Motivations of GRAPE. $\oplus$ represents demand, while $\ominus$ represents supply.

demand, than those predicted by existing models [12], [13]. Such a distribution, if recorded as the input of a prediction model, could lead to biased predictions. Deep learning in the prediction regions, as existing usage [2], [14] might aim to maximize the usage of regions with more usage, which, in terms of prediction accuracy, will affect the different communities, leading to an imbalance in bike sharing networks due to demand and exacerbation of the local issues.

Motivated by the above issues, we propose a novel station-based bike sharing prediction model with **GRAPE**-fusion differentiable pooling with resource and performance fairness. This paper makes the following contributions in addressing the above research gaps:

(i) Analyzing Mobility Patterns (Sec. 2): In this study, we analyze the spatial and pervasive modalities of bike sharing designs. Specifically, we analyze the extensive real-world data analysis of bike sharing temporal characteristics across two metropolitan cities, New York City (NYC) and Chicago. We analyze and quantify the spatial (say, in terms of the common location-wise demand distributions) and temporal (e.g., in terms of the shared repetitive patterns) correlations between different mobility systems from our analysis. Such spatio-temporal correlations motivate our cross-modality graph fusion in GRAPE beyond a single-modality setting. We have also conducted an extensive bike station analysis with the socioeconomic data of NYC and Chicago. We derive their social equity implications in relation to the different modality systems. We observe the unbalanced distributions of bike resources across regions with different socioeconomic features (e.g., ethnicity, education, and income), which motivate our equity-aware prediction modeling.

(ii) Predicting Bike Usage with Cross-modality Graph-Fusion Differentiable Pooling (Secs. 3 and 4): To achieve the cross-modality prediction for the heterogeneous system of target (bike) and auxiliary (taxi) networks, we have designed within GRAPE a novel graph-fusion differentiable pooling framework where the cross-modality spatio-

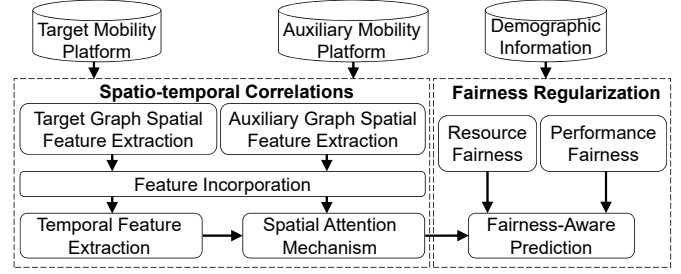


Fig. 2: System overview and information flow of GRAPE.

temporal features of the above two systems are combined throughout the hierarchical space with the shared graph structures. The framework incorporates the hierarchical features of the auxiliary graph into the target graph via a novel convolutional fusion design which captures the spatio-temporal correlations between the target and auxiliary systems. We further augment a spatial attention mechanism with a long short-term memory (LSTM) network to capture the spatio-temporal correlations from the fused hierarchical features, yielding high prediction accuracy.

(iii) Designing Resource and Performance Fairness Metrics for Station-based Bike Sharing Learning (Sec. 5): GRAPE accounts for the socioeconomic information and integrates the fairness metrics with the model regularization to generate equity-aware prediction across the entire bike sharing network. Specifically, we consider fairness in the bike-resource distribution and in the model prediction performance. As for resource, we mitigate the difference in per capita demands across the stations in the advantaged and disadvantaged regions. This way, the predictions can help the bike sharing service providers to adjust future resource allocation equitably and mitigate the unfairness across communities. As to performance, we aim to reduce the prediction-error difference weighted by the populations at the bike sharing stations within advantaged and disadvantaged communities. This way, GRAPE can balance the quality of bike sharing service, say, in terms of prediction accuracy, experienced by the different communities.

(iv) Extensive Experimental Studies (Sec. 6): We have conducted extensive experimental studies upon over 23 million of bike trips and 100 million of taxi trips in total from NYC and Chicago. Our current experimental studies have demonstrated that our proposed model outperforms the other approaches in bike usage prediction, on average by 15.80% for NYC and 50.55% for Chicago in terms of Mean Squared Error and R^2 on average. Furthermore, GRAPE improves the social equity of the bike sharing systems, on average by 32.44% and 24.43% in terms of resource fairness for NYC and Chicago, and 13.36% and 16.52% in terms of performance fairness.

• System Overview. Fig. 2 provides an overview of GRAPE's system framework. We first harvest the mobility data from the target and auxiliary modalities, i.e., the bike and taxi systems in our studies. Then, GRAPE processes and extracts the graph features from the target and auxiliary graphs, and fuses the hierarchical graphs. The embedded spatial features are then integrated for the temporal feature extraction. Finally, GRAPE leverages the socioeconomic data from the U.S. Census for the equity-aware regularization upon the model in order to mitigate the unfairness in

resource prediction and model performance.

• **Societal Implications.** As illustrated in Fig. 1(b), our model output can be fed to the government web portals, enabling responsible city planning [17], digital governance, and civic community engagement. Above and beyond the theoretical analysis and data-driven studies, the insights gained from GRAPE may benefit the following three stakeholder groups: (1) government and city planners in integrating the mobility equity insights to plan bike sharing station distributions; (2) diverse urban communities in benefiting from more inclusive deployment of bike sharing resources/services; and (3) urban computing practitioners and bike sharing service providers in enhancing bike-usage prediction through cross-modality graph-fusion pooling designs. Using our results, the related stakeholders can take various measures, including user survey, feedback crowd-sourcing [18], and public hearings to further investigate the usage and socioeconomic distributions in digital city planning/governance and civic community engagement, as illustrated in Fig. 1(b).

The rest of the paper is organized as follows. We first present the mobility and socioeconomic data analysis in Sec. 2. Given the derived features, we then present the problem definition and the model overview in Sec. 3. We then present the core model of GRAPE in Sec. 4, and designs of the fairness metrics in Sec. 5. We provide the performance evaluations in Sec. 6. We discuss the deployment of GRAPE in Sec. 7, followed by the review of the related work in Sec. 8. We finally conclude the paper in Sec. 9.

2 MOBILITY & SOCIOECONOMIC DATA ANALYSIS

We first detail our real-world datasets of bike sharing, taxi systems, and socioeconomic datasets in Sec. 2.1. We then present the data analysis of the shared spatio-temporal features in Sec. 2.2. After that, we study the distributions of bike resources regarding the socioeconomic characteristics in Sec. 2.3, which further motivates the needs for equity-aware learning and model designs.

2.1 Dataset Overview

To explore the spatio-temporal correlations between bike and taxi systems, we utilize the real-world user trip data of bike sharing platforms of NYC and Chicago. To analyze the socioeconomic features of bike resource distribution, we take into account the socioeconomic data from the U.S. Census Bureau. Each of these datasets is detailed below.

- 1) *User Trip Data of Bike Sharing Systems:* The bike trip dataset describes every single trip including the trip's start/end time, start/end stations, and their longitudes/latitudes. We have collected a total of 23,665,647 trips from two cities: 20,551,697 trips from Citi Bike [19] in NYC during 2019/01/01–2019/12/31 and 3,113,950 trips of Divvy Bike [20] in Chicago during 2019/01/01–2019/09/30.
- 2) *User Trip Data of Taxi Systems:* In this prototype, we use the taxi data as auxiliary information to enhance bike-usage prediction. Similar to bike trip data, the taxi trip dataset describes every single trip including the trip's start/end time, start/end regions (e.g.,

community areas, taxi zones), and the GPS coordinates of the regions' boundaries. We have extracted a total of 100,876,384 trips from the two cities in 2019: 84,399,019 trips from the Yellow Cab in NYC [21] and 16,477,365 trips in Chicago [22].

- 3) *Socioeconomic Data:* To analyze the social equity of bike sharing systems, we have extracted socioeconomic data per census block group for the two cities. We retrieve the social ethnicity distribution, per capita income, and population holding bachelor's degrees in 2018 from U.S. Census Bureau [23]. The dataset contains socioeconomic information in each of the 6,493 census block groups in NYC (i.e., Bronx, Kings, New York, Queens, and Richmond counties) and the 3,993 census block groups in Chicago (Cook county).

We note that all the datasets we studied are open-sourced and all sensitive user information has been sanitized before release by the government portals and service providers, and hence no institutional review board approval is required.

2.2 Shared Spatio-Temporal Mobility Patterns

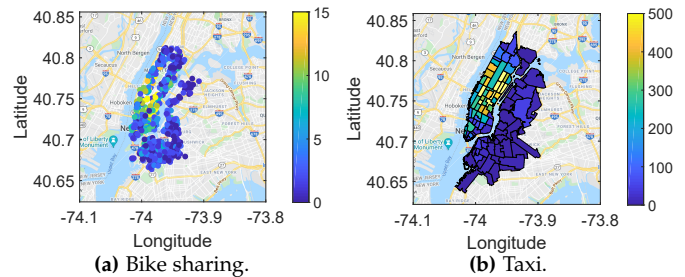


Fig. 3: Pick-up demand heatmaps of bike sharing and taxi systems of NYC on Monday 01-21-2019 from 8:00 to 9:00am. The warmer colors indicate higher demands.

• **Spatial Correlations between Bike Sharing and Taxi Systems.** We first show the similar spatial usage patterns of two mobility systems in NYC in Fig. 3 (with the spatial demand distributions in terms of bike stations and taxi zones that are pre-defined by the service operators). Fig. 3a shows that the bike sharing stations at the center of Manhattan, NYC have higher bike demands than the rest of the city. The taxi system demonstrates spatial distributions similar to the bike sharing system, as illustrated in Fig. 3b. Despite the potentially different commute purposes and mobility preferences of the two mobility systems, such a spatial correlation should be carefully taken into account in our cross-modality modeling.

To further quantify the spatial correlations between the two systems, we divide the city map of NYC into a grid map, each of which is of the size of $1.0 \times 1.0 \text{ km}^2$. We aggregate the total departures or arrivals of each day in all bike stations or taxi regions in each grid. We then flatten the generated grid data into a vector and compute the Pearson's correlation coefficient [24] between the usage vectors of the bike and taxi systems. We can see from the results that the spatial correlation coefficients are mostly around 0.6

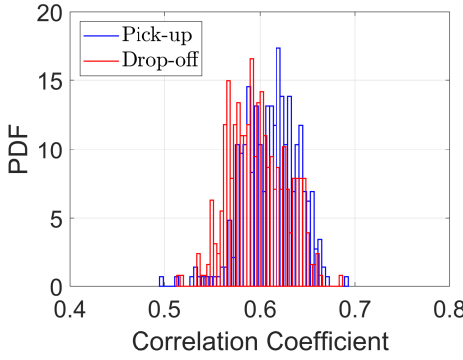


Fig. 4: PDFs of spatial correlations of daily bike and taxi pick-ups/drop-offs in 2019.

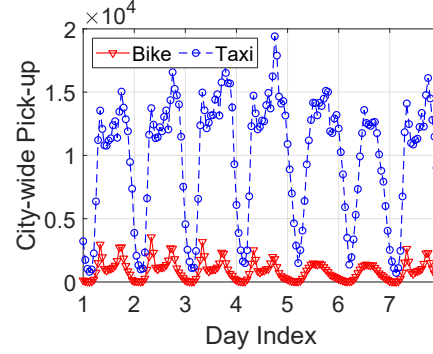


Fig. 5: Hourly bike/taxi departures from 01-08-2019 to 01-14-2019 in NYC.

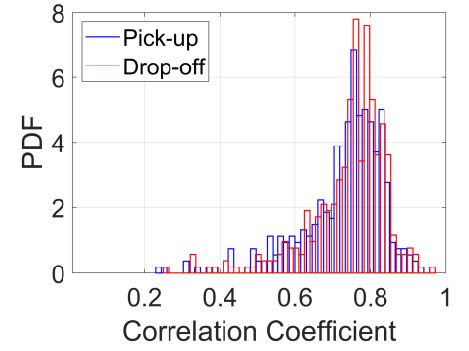


Fig. 6: PDFs of daily temporal correlations in 2019.

(i.e., 0.61 ± 0.03 and 0.60 ± 0.03 for pick-ups and drop-offs, respectively). We have empirically studied different spatial discretization levels, and observed the similar probability density functions (PDFs). Such a shared spatial distribution motivates our designs to characterize the spatial correlations across the modalities in GRAPE.

• **Temporal Correlations between Bike Sharing and Taxi Systems.** Our extensive data analysis also reveals the temporal correlations between bike sharing and taxi systems. Taking NYC as an example, we aggregate the hourly bike/taxi usage of all the stations/regions to generate the hourly bike/taxi usage of the entire city. We first illustrate in Fig. 5 the temporal patterns of city's total bike and taxi demands during a week. We can see that despite the scale and value differences, there exist similar daily mobility patterns of taxi and bike usages due to the repetitive daytime commute. Furthermore, we also observe the similar weekly mobility patterns for both systems; the weekday usage is usually higher than that of the weekends.

In particular, we find the correlation coefficients between the hourly departures/arrivals of the bike and taxi systems in NYC for each week of 2019. We then plot in Fig. 6 the PDFs of the Pearson's correlation coefficients. We can observe generally high temporal correlations between the two systems (i.e., 0.73 ± 0.11 and 0.74 ± 0.10 for pick-ups and drop-offs, respectively). This motivates our cross-modality model designs in GRAPE to capture the temporal correlations of the two systems.

2.3 Analysis of Station-based Bike-Usage Fairness

• **Advantaged and Disadvantaged Communities.** In addition to the cross-modality feature fusion, our work aims at enhancing the fairness in the deployment of mobility systems, particularly focusing on bike sharing. As discussed in Sec. 2.1, we have collected the raw demographic data per census block group. Based on these census tract data, we first define the binary social groups for fairness evaluation following the practices in prior transportation mobility fairness studies [25], [26].

Specifically, we define the advantaged and disadvantaged communities based on socioeconomic attributes in the U.S. Census Tract datasets, and our study focuses on the following three attributes:

(a) *Social Ethnicity:* We consider the geographic regions where the major population is from the social ethnicity other

than Caucasians as the disadvantaged communities, and the regions where the Caucasian population exceeds 50% of the total as the advantaged communities.

(b) *Income Level:* Taking NYC as an example, we illustrate in Fig. 7a the normalized per capita income, which shows the spatial variations of the income levels of different neighborhoods. In this study, we take into account the regions where the per capita/median household annual income is lower than the 30th percentile of that of the city as the disadvantaged communities.

(c) *Education Level:* Our formulation takes in the regions with less than 30% of population holding bachelor's degrees as the disadvantaged communities.

Note that despite our focus upon the above three attributes in this prototype study, GRAPE is general enough to be extended to other socioeconomic attributes [23].

• **Socioeconomic Features & Unbalanced Station Distributions.** Figs. 7b–7d further illustrate the spatial distributions of bike stations located in advantaged and disadvantaged communities in terms of social ethnicity, income, and education. From Figs. 7b–7d, we can see that stations with mostly the advantaged communities (labeled with blue dots) residing nearby are often distributed at the center of Manhattan and the east of Brooklyn. We note that Fig. 3a shows that the major bike resources (stations) are allocated to the mobility demands at the center of Manhattan in NYC. Combining Fig. 3a and Figs. 7b–7d, we can infer that the potential social inequity of deployed bike resources across different communities, where the stations with high demand and, therefore, more bike resources are usually situated in the city regions close to advantaged communities.

Given the above settings and conditions, a conventional deep learning model [2], [27] trained with unbalanced station distributions and historical bike usage may create biased bike-usage prediction favoring these regions with advantaged communities. The prior practices might escalate the divergence in the future resource allocation as well as service quality. This motivates our designs in the fairness metrics regarding bike resources and prediction errors as *regularizers* in our cross-modality model.

3 PROBLEM & MODEL OVERVIEW

Given the above data analysis, we first define and formalize the core problem in Sec. 3.1, and overview our GRAPE design in Sec. 3.2.

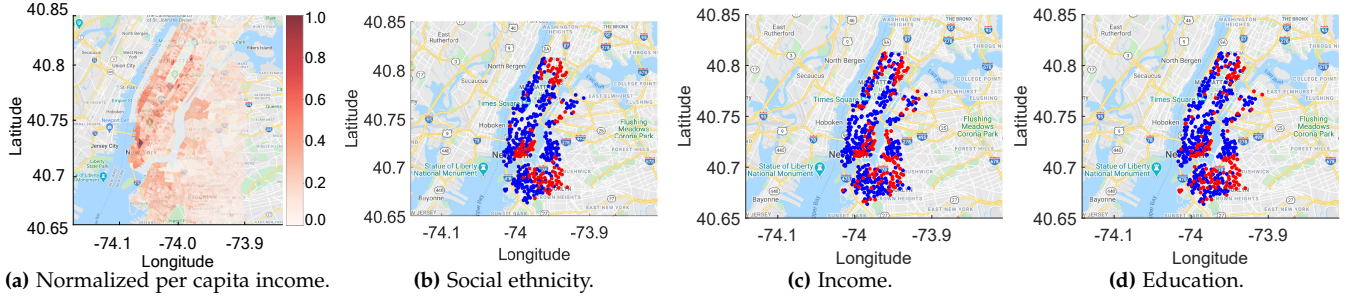


Fig. 7: (a) Normalized per capita income. (b)–(d) Bike station distributions of NYC. The blue dots represent the stations located in the areas where the advantaged community predominates, and the red dots otherwise.

3.1 Problem Definition

• **Bike Sharing Usage Prediction.** Considering a bike sharing network of N stations as the target system, denoted as “tar”, each station i has two features, $\mathbf{z}_{\text{tar},k}^{(d)}[i]$ and $\mathbf{z}_{\text{tar},k}^{(r)}[i]$, i.e., the aggregate numbers of departures (d) and arrivals (r) from/at station i within a time interval k . Taking in a time window of $\mathcal{T}$ such intervals, GRAPE forms

$$\mathbf{z}_{\text{tar}}^{(d)} = [\mathbf{z}_{\text{tar}}^{(d)}[1], \dots, \mathbf{z}_{\text{tar}}^{(d)}[N]],$$

and

$$\mathbf{z}_{\text{tar}}^{(r)} = [\mathbf{z}_{\text{tar}}^{(r)}[1], \dots, \mathbf{z}_{\text{tar}}^{(r)}[N]],$$

where $\mathbf{z}_{\text{tar}}^{(d)}[i] = [\mathbf{z}_{\text{tar},t-\mathcal{T}+1}^{(d)}[i], \mathbf{z}_{\text{tar},t-\mathcal{T}+2}^{(d)}[i], \dots, \mathbf{z}_{\text{tar},t}^{(d)}[i]]$ and $\mathbf{z}_{\text{tar}}^{(r)}[i] = [\mathbf{z}_{\text{tar},t-\mathcal{T}+1}^{(r)}[i], \mathbf{z}_{\text{tar},t-\mathcal{T}+2}^{(r)}[i], \dots, \mathbf{z}_{\text{tar},t}^{(r)}[i]]$ for $i \in [1, \dots, N]$, representing the $\mathcal{T}$ historical bike shares or pick-ups, and arrivals or drop-offs from i by the present time interval t .

Similarly to $\mathbf{z}_{\text{tar}}^{(d)}, \mathbf{z}_{\text{tar}}^{(r)} \in \mathbb{R}^{N \times \mathcal{T}}$, taking into account the taxi system as the auxiliary system, denoted as “aux”, we find $\mathbf{z}_{\text{aux}}^{(d)}, \mathbf{z}_{\text{aux}}^{(r)} \in \mathbb{R}^{N' \times \mathcal{T}}$ for the taxi departures d of N' taxi regions.

Using the above setting, the bike sharing usage prediction problem of GRAPE is defined as follows. Given the bike/taxi usage of each station/region of previous intervals, $\mathbf{z}_{\text{tar}}^{(d)}, \mathbf{z}_{\text{tar}}^{(r)}, \mathbf{z}_{\text{aux}}^{(d)}$, and $\mathbf{z}_{\text{aux}}^{(r)}$, we aim to target bike departures and arrivals beyond the p interval t ,

$$\hat{\mathbf{z}}_{\text{tar}}^{(d)}[i] = [\hat{\mathbf{z}}_{\text{tar},t+1}^{(d)}[i], \hat{\mathbf{z}}_{\text{tar},t+2}^{(d)}[i], \dots, \hat{\mathbf{z}}_{\text{tar},t+\mathcal{F}}^{(d)}[i]]$$

and

$$\hat{\mathbf{z}}_{\text{tar}}^{(r)}[i] = [\hat{\mathbf{z}}_{\text{tar},t+1}^{(r)}[i], \hat{\mathbf{z}}_{\text{tar},t+2}^{(r)}[i], \dots, \hat{\mathbf{z}}_{\text{tar},t+\mathcal{F}}^{(r)}[i]]$$

at each station i with the horizon of future $\mathcal{F}$ time intervals.

• **Bike Sharing System Equity.** To further achieve the equity-aware prediction, we will take into account resource distribution and model performance in regularizing the GRAPE model:

- **Resource Fairness:** For resource distribution, our regularization mitigates the unbalanced bike resource distribution between stations in the advantaged and disadvantaged communities to reduce the unfairness of resource allocation of the bike sharing system. Specifically, we reduce the difference of the per capita demand predictions between the stations that are near the advantaged communities $i^+ \in [1, \dots, N^+]$ and those near the disadvantaged communities, $i^- \in [1, \dots, N^-]$,

$[1, \dots, N^-]$, $\frac{\sum_{i^+}^{N^+} \hat{\mathbf{z}}_{\text{tar},k}^{(d)}[i^+]}{\sum_{i^+}^{N^+} D_{i^+}}$ and $\frac{\sum_{i^-}^{N^-} \hat{\mathbf{z}}_{\text{tar},k}^{(d)}[i^-]}{\sum_{i^-}^{N^-} D_{i^-}}$, where N^+ (N^-) is the total number of stations with major advantaged (disadvantaged) communities nearby ($N^+ + N^- = N$), and D_{i^+} and D_{i^-} denote their respective ratios of the population of the census block group to the total population of the city where the stations i^+ and i^- are located.

- **Performance Fairness:** To evaluate the performance fairness, our regularization reduces the difference in mean squared errors (MSEs) of predictions which are weighted by the populations of advantaged and disadvantaged communities, D_{i^+} and D_{i^-} , where the stations i^+ and i^- are located. Here we incorporate the population as the weights in order to ensure a reasonable model prediction accuracy (and

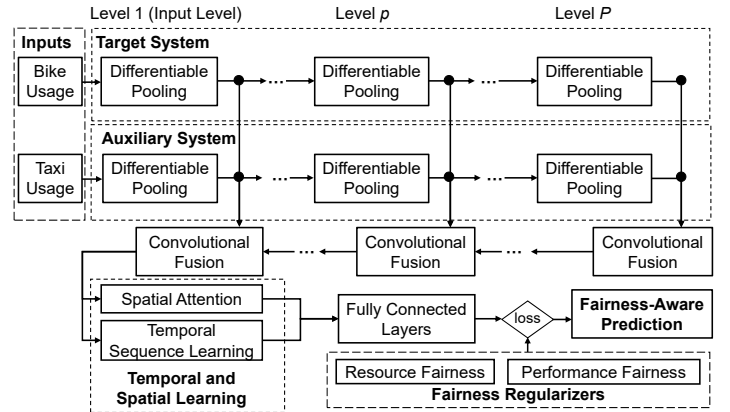


Fig. 8: Overview of the core structure of GRAPE.

3.2 Overview of Core Model and Hierarchical Graphs

• **Model Overview.** To jointly approach the prediction and equity problems, we have designed GRAPE, an equity-aware bike-usage prediction system, and its core structure is illustrated in Fig. 8.

Specifically, we first formulate the bike sharing network as a target graph, G_{tar} , and the taxi system as an auxiliary graph, G_{aux} , where G_{tar} and G_{aux} have different numbers of nodes. Given the thus-constructed target and auxiliary graphs from the input level, we then extract their hierarchical spatial features using the separate differential pooling mechanisms (see Sec. 4.1) from the levels 1 to P . This way,

at each hierarchical level (level p) the nodes in the target and auxiliary graphs are assigned, respectively, to the same number of nodes, enabling fusion of hierarchical spatial features of the auxiliary graph into those of target graph.

The extracted features of the auxiliary graph are then incorporated into the target graph at hierarchical levels via a convolutional fusion mechanism (see Sec. 4.1) in a backward direction from the level P to the input level. We further leverage temporal and spatial learning module, which consists of temporal sequence learning and spatial attention (see Sec. 4.2), to capture the temporal and spatial features of the updated embeddings of the target graph.

To mitigate the social inequity of the target (bike) systems in terms of resource and performance, we have designed two fairness metrics (Sec. 5), i.e., fairness in resource allocation and model performance, as additional objective functions for GRAPE, leading to equity-aware prediction.

• **Target and Auxiliary Graphs.** For the target graph, G_{tar} , each node represents a bike station with bike departure and arrival volumes as its two features, and there are N nodes considering a system with N bike stations. We consider an edge between any two nodes representing users can commute between any two stations in the city. We note that in our experimental studies, as the pick-up and drop-off locations of the NYC taxis are provided at the zone or region level instead of fine-grained coordinates, we formulate the bike station network as a graph at each node in the graph represent such a region including taxi departure and arrival volumes as its two features. For a taxi system with N' taxi zones or communities, the auxiliary graph, denoted as G_{aux} , contains N' nodes. Similarly, we construct an edge between any two nodes in G_{aux} to represent the commutes between any two regions in the city.

The weight of each edge in G_{tar} or G_{aux} represents the correlation between two city locations. In this work, we adopt the inverse of *geo-distances* between locations as their correlations to formalize each element inside the adjacency matrix of the constructed graph. The adjacency matrix at the input level for G_{tar} , denoted as $\mathbf{A}_{\text{tar}}^{(1)} \in \mathbb{R}^{N \times N}$, is formulated as

$$\mathbf{A}_{\text{tar}}^{(1)}[i, j] = \mathbf{A}_{\text{tar}}^{(1)}[j, i] = \frac{1}{g_{\text{tar}}(i, j)}, \quad (3)$$

where $g_{\text{tar}}(i, j)$ denotes the geo-distance (in km) between two locations i and j . Similarly we can find the weight in the auxiliary graph $\mathbf{A}_{\text{aux}}^{(1)}[i, j]$ based on $g_{\text{aux}}(i, j)$, where $\mathbf{A}_{\text{aux}}^{(1)} \in \mathbb{R}^{N' \times N'}$. We set the diagonal elements of the adjacency matrix $\mathbf{A}_{\text{tar}}^{(1)}[i, i] = 0$ and $\mathbf{A}_{\text{aux}}^{(1)}[i, i] = 0$. In practice, for the taxi graph, the regions of departures/arrivals are usually in polygons. For simplicity, we use the center points of their bounding boxes to calculate the geo-distances between each two regions as their correlations.

4 CORE MODEL DESIGN

We present in the following the core model designs of GRAPE, i.e., the hierarchical graph learning design in Sec. 4.1, and temporal and spatial learning design in Sec. 4.2.

4.1 Hierarchical Graph Learning Designs

Our hierarchical graph learning for cross-modality fusion includes the following two major components, i.e., (1) the graph differentiable pooling, and (2) the convolutional fusion of graph embeddings, which will be discussed in details below.

(1) **Graph Differentiable Pooling.** From Sec. 2.2, we learn that the mobility patterns of the target and auxiliary systems have latent spatio-temporal characteristics shared at the city regions. The graph differentiable pooling design of GRAPE aims to merge the knowledge of the auxiliary system into the target system at the hierarchical level to improve the target graph representation and the subsequent prediction accuracy. Specifically, we first design the hierarchical graph embedding to achieve such knowledge fusion across the target and auxiliary mobility systems, via a graph differentiable pooling mechanism [28]. Taking the target graph as

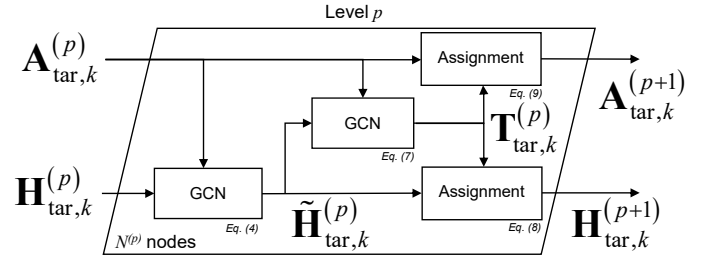


Fig. 9: An example of graph differentiable pooling mechanism at the p -th hierarchical level of the target graph.

(a) *Graph Convolution.* As illustrated in Fig. 9, for each level, p , we form a total of $N^{(p)}$ nodes by assigning the $N^{(p-1)}$ nodes from the previous level $p-1$. At each level for the time interval k ($k \in \{t-T+1, \dots, t\}$), we first perform a GCN operation [29] on the input embeddings, $\mathbf{H}_{\text{tar},k}^{(p)}$, and the adjacency matrix $\mathbf{A}_{\text{tar},k}^{(p)}$ assigned from the preceding level $p-1$ to process and capture the spatial correlations between $N^{(p)}$ nodes, and generate the GCN embeddings at this level, $\tilde{\mathbf{H}}_{\text{tar},k}^{(p)}$, i.e.,

$$\tilde{\mathbf{H}}_{\text{tar},k}^{(p)} = \text{GCN} \left(\mathbf{A}_{\text{tar},k}^{(p)}, \mathbf{H}_{\text{tar},k}^{(p)} \right) \in \mathbb{R}^{N^{(p)} \times \tilde{f}_{\text{tar}}^{(p)}}, \quad (4)$$

where the superscript p denotes the p -th hierarchical level, and $\tilde{f}_{\text{tar}}^{(p)}$ represents the GCN embedding size. Note that level $p=1$ denotes the input level, and we let $\mathbf{H}_{\text{tar},k}^{(1)}$ be the bike pick-ups/drop-offs at the time interval k , and $\mathbf{A}_{\text{tar},t-T+1}^{(1)} = \dots = \mathbf{A}_{\text{tar},k}^{(1)} = \dots = \mathbf{A}_{\text{tar},t}^{(1)} = \mathbf{A}_{\text{tar}}^{(1)}$ be the adjacency matrices at the input level (Eq. (3)).

Within the $\text{GCN}(\cdot)$ operation, each element of the adjacency matrix $\mathbf{A}_{\text{tar},k}^{(p)}[i, j]$ is first normalized, i.e.,

$$\tilde{\mathbf{A}}_{\text{tar},k}^{(p)} = \mathbf{D}^{-\frac{1}{2}} \left(\mathbf{A}_{\text{tar},k}^{(p)} + \mathbf{I} \right) \mathbf{D}^{\frac{1}{2}}, \quad (5)$$

where $\mathbf{D} \in \mathbb{R}^{N^{(p)} \times N^{(p)}}$ is the degree matrix of $\mathbf{A}_{\text{tar},k}^{(p)}$ and $\mathbf{I} \in \mathbb{R}^{N^{(p)} \times N^{(p)}}$ is an identity matrix. Then the convolutional operation is given by

$$\tilde{\mathbf{H}}_{\text{tar},k}^{(p)} = \tilde{\mathbf{A}}_{\text{tar},k}^{(p)} \mathbf{H}_{\text{tar},k}^{(p)} \mathbf{W}_{\text{tar}}^{(p)}, \quad (6)$$

where $\mathbf{W}_{\text{tar}}^{(p)} \in \mathbb{R}^{f_{\text{tar}}^{(p)} \times \tilde{f}_{\text{tar}}^{(p)}}$ represent.

(b) *Node Assignment*. We have ages the $\tilde{\mathbf{H}}_{\text{tar},k}^{(p)}$ and the adjacency the transition matrix, $\mathbf{T}_{\text{tar},k}^{(p)} \in \mathbb{R}^{\Lambda}$ matrix $\mathbf{T}_{\text{tar},k}^{(p)}$ will be used for assi the $(p+1)$ -th level, and the resu again become the input of the su Such node assignment mechanis spatio-temporal knowledge share enabling the embedding fusion.

We detail our node assignment ing. To assign the nodes at the p -t $(p+1)$ -th level, we generate the tr GCN. Specifically, the GCN takes in for the time interval k :

$$\mathbf{T}_{\text{tar},k}^{(p)} = \text{softmax} \left(\text{GCN} \right)$$

where $\mathbf{T}_{\text{tar},k}^{(p)} \in \mathbb{R}^{N^{(p)} \times N^{(p+1)}}$ is t the levels p to $p+1$, and $\mathbf{A}_{\text{tar},k}^{(p)} \in \mathbb{R}$ matrix at the p -th level.

We let the output dimension the number of nodes at the $(p+1)$ -th level, $N^{(p+1)}$. We use the softmax function to normalize each row of the transition matrix, and hence the transpose of the transition matrix, $(\mathbf{T}_{\text{tar},k}^{(p)})^T \in \mathbb{R}^{N^{(p+1)} \times N^{(p)}}$, represents the *probability* of assigning the nodes at the p -th level to each one of the nodes at the $(p+1)$ -th level. This probability encodes which target nodes in a graph belong to the same geospatial region (a node at the next level) based on the spatial correlations between the nodes (further visualization and interpretation can be referred to Sec. 6.2).

Then, we obtain the input embeddings of the $(p+1)$ -th level, i.e.,

$$\mathbf{H}_{\text{tar},k}^{(p+1)} = (\mathbf{T}_{\text{tar},k}^{(p)})^T \tilde{\mathbf{H}}_{\text{tar},k}^{(p)}, \quad (8)$$

where $\mathbf{H}_{\text{tar},k}^{(p+1)} \in \mathbb{R}^{N^{(p+1)} \times f_{\text{tar}}^{(p+1)}}$ represents the input embeddings of the GCN at the $(p+1)$ -th level, and $f_{\text{tar}}^{(p+1)} = \tilde{f}_{\text{tar}}^{(p)}$. Similarly, taking into account $\mathbf{T}_{\text{tar},k}^{(p)}$ as the probability of node assignment, the adjacency matrix at the $(p+1)$ -th level becomes

$$\mathbf{A}_{\text{tar},k}^{(p+1)} = (\mathbf{T}_{\text{tar},k}^{(p)})^T \mathbf{A}_{\text{tar},k}^{(p)} \mathbf{T}_{\text{tar},k}^{(p)}, \quad (9)$$

where $\mathbf{A}_{\text{tar},k}^{(p+1)} \in \mathbb{R}^{N^{(p+1)} \times N^{(p+1)}}$ is the adjacency matrix at the $(p+1)$ -th level.

Following the same manner, we feed $\mathbf{A}_{\text{aux},k}^{(p)}$ and $\mathbf{H}_{\text{aux},k}^{(p)}$ to the graph differentiable pooling module of the auxiliary graph, and obtain the $\mathbf{T}_{\text{aux},k}^{(p)}$, $\mathbf{H}_{\text{aux},k}^{(p+1)}$, and $\mathbf{A}_{\text{aux},k}^{(p+1)}$ at the p -th hierarchical level of the auxiliary graph.

(2) Convolutional Fusion of Graph Embeddings. Given the embeddings [30] of the target and auxiliary graphs at each hierarchical level for the past $\mathcal{T}$ time intervals, i.e., $\mathbf{H}_{\text{tar}}^{(p)} \in \mathbb{R}^{\mathcal{T} \times N^{(p)} \times f_{\text{tar}}^{(p)}} = [\mathbf{H}_{\text{tar},t-\mathcal{T}+1}^{(p)}, \dots, \mathbf{H}_{\text{tar},t}^{(p)}]$ and $\mathbf{H}_{\text{aux}}^{(p)} \in \mathbb{R}^{\mathcal{T} \times N^{(p)} \times f_{\text{aux}}^{(p)}} = [\mathbf{H}_{\text{aux},t-\mathcal{T}+1}^{(p)}, \dots, \mathbf{H}_{\text{aux},t}^{(p)}]$ ($f_{\text{aux}}^{(p)}$ represents the channel size for the auxiliary graph at the level p), we fuse them in a backward manner from the levels P to 1.

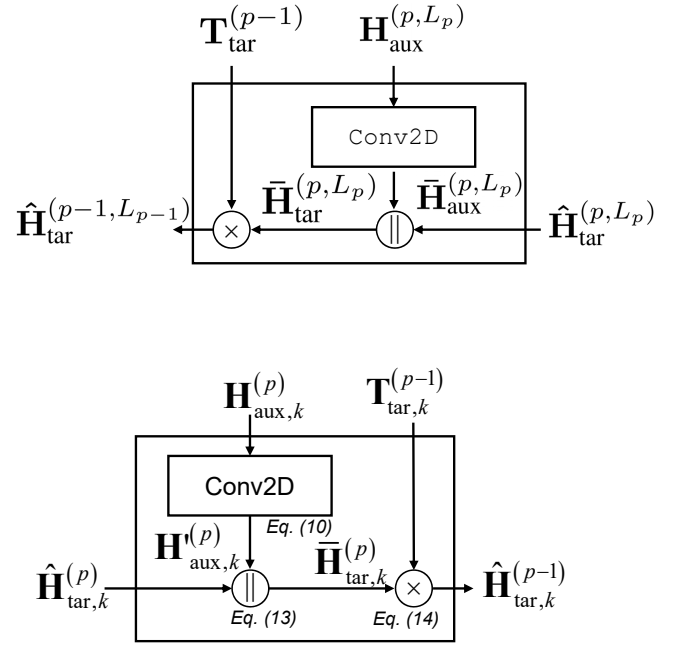


Fig. 10: Illustration of convolutional fusion mechanism across the levels p and $p-1$. The mechanism takes in the embeddings of $\mathbf{H}_{\text{aux}}^{(p)}$ and $\hat{\mathbf{H}}_{\text{tar}}^{(p)}$, and the probability of assignment, $\mathbf{T}_{\text{tar},k}^{(p)}$, and returns $\hat{\mathbf{H}}_{\text{tar},k}^{(p-1)}$ for the preceding layer. The operator $\parallel$ denotes concatenation operation and $\times$ denotes matrix multiplication.

Specifically, as illustrated in Fig. 10, for historical embeddings at the p -th hierarchical level for auxiliary graph, $\mathbf{H}_{\text{aux}}^{(p)} \in \mathbb{R}^{\mathcal{T} \times N^{(p)} \times f_{\text{aux}}^{(p)}}$, we adopt a 2-D convolutional neural network (Conv2D) with zero padding to learn their weekly patterns and obtain the updated embeddings of the auxiliary graph, i.e.,

$$\mathbf{H}'_{\text{aux}}^{(p)} = \text{Conv2D} \left(\mathbf{H}_{\text{aux}}^{(p)} \right), \quad (10)$$

where $\mathbf{H}'_{\text{aux}}^{(p)} \in \mathbb{R}^{\mathcal{T} \times N^{(p)} \times f_{\text{aux}}^{(p)'}}$.

Then, we fuse the updated auxiliary embeddings with the target embeddings, $\mathbf{H}_{\text{tar}}^{(p)} \in \mathbb{R}^{\mathcal{T} \times N^{(p)} \times f_{\text{tar}}^{(p)}}$, at each hierarchical level via the concatenation operation. We use the concatenation operation instead of summation fusion to increase the flexibility of the model as the summation requires the same matrix dimension making the model less robust. Specifically, at the level P (at the top of the hierarchical structure), the updated auxiliary embeddings $\mathbf{H}'_{\text{aux}}^{(P)}$ are concatenated with the original target graph embeddings $\mathbf{H}_{\text{tar}}^{(P)}$, i.e.,

$$\bar{\mathbf{H}}_{\text{tar}}^{(P)} = \mathbf{H}_{\text{tar}}^{(P)} \parallel \mathbf{H}'_{\text{aux}}^{(P)}, \quad (11)$$

where $\bar{\mathbf{H}}_{\text{tar}}^{(P)} \in \mathbb{R}^{\mathcal{T} \times N^{(P)} \times \bar{f}_{\text{tar}}^{(P)}}$ and $\bar{f}_{\text{tar}}^{(P)} = f_{\text{aux}}^{(P)} + f_{\text{tar}}^{(P)}$. We update the embeddings of the $(P-1)$ -th level by projecting $\bar{\mathbf{H}}_{\text{tar}}^{(P)}$ back to the level $(P-1)$ at each time interval [31], i.e.,

$$\hat{\mathbf{H}}_{\text{tar},k}^{(P-1)} = \mathbf{T}_{\text{tar},k}^{(P-1)} \bar{\mathbf{H}}_{\text{tar},k}^{(P)}, \quad (12)$$

and we have $\hat{\mathbf{H}}_{\text{tar}}^{(P-1)} = [\hat{\mathbf{H}}_{\text{tar},t-\mathcal{T}+1}^{(P-1)}, \dots, \hat{\mathbf{H}}_{\text{tar},t}^{(P-1)}]$ for all the time intervals.

For the subsequent layers other than the P -th level, we update embeddings of the target graph by the concatenating the projection of the updated target graph embeddings at

the level p for all time intervals, $\hat{\mathbf{H}}_{\text{tar}}^{(p)}$, and the convolution of the auxiliary graph embeddings at the level p , $\mathbf{H}'_{\text{aux}}^{(p)}$, i.e.,

$$\bar{\mathbf{H}}_{\text{tar}}^{(p)} = \hat{\mathbf{H}}_{\text{tar}}^{(p)} \parallel \mathbf{H}'_{\text{aux}}^{(p)}. \quad (13)$$

Then we project $\bar{\mathbf{H}}_{\text{tar}}^{(p)}$ back to the level $(p-1)$ at each time interval, i.e.,

$$\hat{\mathbf{H}}_{\text{tar},k}^{(p-1)} = \mathbf{T}_{\text{tar},k}^{(p-1)} \bar{\mathbf{H}}_{\text{tar},k}^{(p)}. \quad (14)$$

Following the above fusion and projection processes, we obtain the updated embeddings of the target graph at the input level, $\bar{\mathbf{H}}_{\text{tar}}^{(1)} \in \mathbb{R}^{\mathcal{T} \times N \times \bar{f}_{\text{tar}}^{(1)}}$, which are fed into the subsequent LSTM module for capturing the temporal and sequential characteristics.

4.2 Temporal and Spatial Learning

We further design the temporal sequence learning and spatial attention learning modules to extract the temporal and spatial features from the generated graph embeddings.

• **Temporal Sequence Learning.** The updated embeddings of the station level, $\bar{\mathbf{H}}_{\text{tar}}^{(1)} \in \mathbb{R}^{\mathcal{T} \times N \times \bar{f}_{\text{tar}}^{(1)}}$, is reshaped into $\bar{\mathbf{H}}_{\text{tar}}^{(1)} \in \mathbb{R}^{\mathcal{T} \times N \times \bar{f}_{\text{tar}}^{(1)}}$ and fed to an LSTM module. We obtain the last hidden state (the present time interval t) $\mathbf{h}_t = \text{LSTM}(\bar{\mathbf{H}}_{\text{tar}}^{(1)})$ where $\mathbf{h}_t \in \mathbb{R}^h$ and h is the number of hidden states.

• **Spatial Attention Learning.** Our spatial attention further captures the multi-level spatial correlations across the time series of different bike stations. Specifically, we first leverage a 1-D convolutional layer (Conv1D) with kernel size $\mathcal{T}$ upon the time dimension (the first dimension of $\bar{\mathbf{H}}_{\text{tar}}^{(1)}$) to encode the temporal patterns of each bike stations, i.e.,

$$\bar{\mathbf{H}}^{\text{att}} = \text{Conv1D}(\bar{\mathbf{H}}_{\text{tar}}^{(1)}), \quad (15)$$

where $\bar{\mathbf{H}}^{\text{att}} \in \mathbb{R}^{N \times f^{\text{att}}}$ represents the resulting station embeddings, and f^{att} is the channel size of Conv1D. This way, we can directly achieve the internal representation of the time series data.

Then, we multiply $\bar{\mathbf{H}}^{\text{att}}$ with the LSTM output hidden state $\mathbf{h}_t$ to compute attention scores for all stations, $\alpha = [\alpha_1, \dots, \alpha_i, \dots, \alpha_N]$, i.e.,

$$\alpha = \text{softmax}(\bar{\mathbf{H}}^{\text{att}} \mathbf{W}^{\text{att}} \mathbf{h}_t), \quad (16)$$

where $\mathbf{W}^{\text{att}} \in \mathbb{R}^{f^{\text{att}} \times h}$ is a trainable parameter. We then sum up the station embeddings based on their attention scores and generate the attention hidden state, $\mathbf{h}_{\text{att}} \in \mathbb{R}^{f^{\text{att}}}$, i.e.,

$$\mathbf{h}_{\text{att}} = \sum_i \alpha_i \bar{\mathbf{H}}^{\text{att}}[i]. \quad (17)$$

Then, we update final hidden state by a linear (dense) layer of the concatenation of $\mathbf{h}_{\text{att}}$ and $\mathbf{h}_t$, i.e.,

$$\bar{\mathbf{h}}_t = \mathbf{W}^{\text{hidden}} \left(\mathbf{h}_{\text{att}} \parallel \mathbf{h}_t \right) + \mathbf{b}^{\text{hidden}}, \quad (18)$$

where $\mathbf{W}^{\text{hidden}} \in \mathbb{R}^{h \times (f^{\text{att}} + h)}$ and $\mathbf{b}^{\text{hidden}} \in \mathbb{R}^h$ are trainable parameters.

A fully connected layer with a ReLU activation function takes the updated last hidden state, $\bar{\mathbf{h}}_t$, to generate predictions of bike departures and arrivals simultaneously. The

final prediction of the future $\mathcal{F}$ time intervals for all N bike sharing stations, $\hat{\mathbf{z}} \in \mathbb{R}^{2\mathcal{F}N}$, is formulated by

$$\hat{\mathbf{z}} = \text{ReLU}(\mathbf{W}^{\text{out}} \bar{\mathbf{h}}_t + \mathbf{b}^{\text{out}}), \quad (19)$$

where $\mathbf{W}^{\text{out}} \in \mathbb{R}^{2\mathcal{F}N \times h}$ and $\mathbf{b}^{\text{out}} \in \mathbb{R}^{2\mathcal{F}N}$. We reshape $\hat{\mathbf{z}}$ into $[\hat{\mathbf{z}}_{\text{tar}}^{(d)}, \hat{\mathbf{z}}_{\text{tar}}^{(r)}]$ (see Eqs. (1) and (2)) as the final outputs.

5 EQUITY-AWARE OBJECTIVE DESIGNS

Based on the aforementioned model designs, we further present in Sec. 5.1 the objective designs of GRAPE in prediction accuracy, as well as the fairness metrics for the equity in terms of resource allocation and performance in Sec. 5.2.

5.1 Prediction Accuracy Metrics

To train GRAPE for predicting the bike usage at different stations, our objectives in terms of prediction accuracy are two-fold.

• **Prediction Accuracy.** First, for *usage prediction accuracy*, GRAPE minimizes the Mean Square Error (MSE), i.e.,

$$\begin{aligned} \mathcal{L}_{\text{MSE}} = & \frac{1}{2N\mathcal{F}} \sum_{i=1}^N \sum_{k=t+1}^{t+\mathcal{F}} \left(\hat{\mathbf{z}}_{\text{tar},k}^{(d)}[i] - \mathbf{z}_{\text{tar},k}^{(d)}[i] \right)^2 \\ & + \frac{1}{2N\mathcal{F}} \sum_{i=1}^N \sum_{k=t+1}^{t+\mathcal{F}} \left(\hat{\mathbf{z}}_{\text{tar},k}^{(r)}[i] - \mathbf{z}_{\text{tar},k}^{(r)}[i] \right)^2, \end{aligned} \quad (20)$$

where $\hat{\mathbf{z}}_{\text{tar},k}^{(d)}$ and $\hat{\mathbf{z}}_{\text{tar},k}^{(r)}$ are the predicted and ground-truth departures, while $\hat{\mathbf{z}}_{\text{tar},k}^{(r)}$ and $\mathbf{z}_{\text{tar},k}^{(r)}$ are the predicted and ground-truth arrivals.

• **Node Assignment.** However, minimizing MSE does not necessarily benefit the hierarchical graph learning due to the non-convexity of the graph optimization [28]. To address this issue, we introduce an auxiliary objective function, i.e., the *node assignment* loss. We leverage this loss to ensure that the spatio-temporal embeddings of the auxiliary graph at the hierarchical levels will be assigned to the correct nodes inside the target graph. It is formally given by

$$\begin{aligned} \mathcal{L}_{\text{tra}} = & \sum_{k=t+1}^{t+\mathcal{F}} \sum_{m \in [\text{tar}, \text{aux}]} \sum_{p=1}^{P-1} \frac{\left\| \mathbf{H}_{m,k}^{(p)} - \mathbf{T}_{m,k}^{(p)} \left(\mathbf{T}_{m,k}^{(p)} \right)^{\top} \mathbf{H}_{m,k}^{(p)} \right\|_F^2}{N^{(p)}} \\ & + \sum_{k=t+1}^{t+\mathcal{F}} \sum_{p=2}^P \left\| \mathbf{A}_{\text{tar},k}^{(p)} - \mathbf{A}_{\text{aux},k}^{(p)} \right\|_F. \end{aligned} \quad (21)$$

The Frobenius norms in Eq. (21) help capture the differences of the input embeddings (matrices) [32]. Specifically, the *first* squared Frobenius norm, denoted as $\|\cdot\|_F$, captures the difference between the original embeddings from the target and auxiliary graphs, $\mathbf{H}_{\text{tar},k}^{(p)}$ and $\mathbf{H}_{\text{aux},k}^{(p)}$, and the embeddings after pooling and back projection for all levels of the target/auxiliary graphs ($p \in \{1, \dots, P-1\}$). The *second* Frobenius norm denotes the difference between the adjacency matrices of the target and auxiliary graphs, $\mathbf{A}_{\text{tar}}^{(p)}$ and $\mathbf{A}_{\text{aux}}^{(p)}$, at each hierarchical level.

In other words, Eq. (21) ensures that (i) the embeddings after pooling and backward projection should be close to the original input level; (ii) the adjacency matrices of the target and auxiliary graphs, representing the node correlations, should be enforced to be the same.

5.2 Designs of Fairness Metrics

In order to reduce the social inequity of our target systems, we propose two fairness metrics in our model studies, i.e., the fairness in *resource distribution* and the fairness in *model performance weighted by population*.

• **Resource Distribution Fairness.** This metric is designed to mitigate the difference between the resources across the advantaged and the disadvantaged communities [33]. We define the fairness metrics in the bike station resource distribution as difference in the per capita predicted demand (pick-up) between the stations in advantaged and the disadvantaged communities (in terms of a certain socioeconomic attribute such as social ethnicity). Specifically, we have the loss as

$$\mathcal{L}_{\text{res}} = \frac{1}{\mathcal{F}} \sum_{k=t+1}^{t+F} \frac{1}{\sum_i^N \mathbf{z}_{\text{tar},k}^{(d)}[i]} \left| \frac{\sum_{i^+}^{N^+} \hat{\mathbf{z}}_{\text{tar},k}^{(d)}[i^+]}{\sum_{i^+}^{N^+} D_{i^+}} - \frac{\sum_{i^-}^{N^-} \hat{\mathbf{z}}_{\text{tar},k}^{(d)}[i^-]}{\sum_{i^-}^{N^-} D_{i^-}} \right|, \quad (22)$$

where $i^+ \in \{1, \dots, N^+\}$ is the index of a station in advantaged communities and $i^- \in \{1, \dots, N^-\}$ represents that of a station in disadvantaged communities. $N^+ + N^- = N$, and D_i is the ratio of the population of the census block group where the station i is located to the total population of the city.

In other words, $\mathcal{L}_{\text{res}}$ decreases if the difference in the per capita demand drops across the stations near the advantaged and the disadvantaged communities, introducing a fairer distribution of bike mobility resources. As here we focus on the station-based bike sharing, we leverage the exact census block groups provided by U.S. Census Tract where the stations are situated for population calculation instead of spatial partitioning and estimation, which may not necessarily reflect the populations served [33].

• **Performance Fairness Adjusted with Population Ratio.** This metric is used to measure the difference of prediction errors weighted by the populations for stations with the major advantaged and disadvantaged communities nearby (in terms of a certain socioeconomic attribute such as social ethnicity). We first find the two MSEs of the stations where the advantaged and disadvantaged communities predominate given the predictions of GRAPE, i.e.,

$$\begin{aligned} \mathcal{E}_{\text{per}^+} &= \frac{1}{2\mathcal{F}N^+} \sum_{q \in \{d,r\}} \sum_{i=1}^{N^+} \frac{D_{i^+}}{\mathbb{E}[D_{i^+}]} \sum_{k=t+1}^{t+F} \left(\hat{\mathbf{z}}_{\text{tar},k}^{(q)}[i^+] - \mathbf{z}_{\text{tar},k}^{(q)}[i^+] \right)^2, \\ \mathcal{E}_{\text{per}^-} &= \frac{1}{2\mathcal{F}N^-} \sum_{q \in \{d,r\}} \sum_{i=1}^{N^-} \frac{D_{i^-}}{\mathbb{E}[D_{i^-}]} \sum_{k=t+1}^{t+F} \left(\hat{\mathbf{z}}_{\text{tar},k}^{(q)}[i^-] - \mathbf{z}_{\text{tar},k}^{(q)}[i^-] \right)^2, \end{aligned} \quad (23)$$

where we adjust the population ratio, denoted as D_i , by its mean over all regions, denoted as $\mathbb{E}[D_i]$, so that $\mathcal{E}_{\text{per}^+}$ and $\mathcal{E}_{\text{per}^-}$ are in the same magnitude of $\mathcal{L}_{\text{MSE}}$ to mitigate the impacts from the over-populated regions. Then, the fairness score in the weighted MSE is given by

$$\mathcal{L}_{\text{per}} = (\mathcal{E}_{\text{per}^+} - \mathcal{E}_{\text{per}^-})^2. \quad (24)$$

In other words, the decrease in $\mathcal{L}_{\text{per}}$ results in a more balanced performance of usage predictions between the stations in the advantaged and the disadvantaged communities.

In summary, we jointly take in the prediction accuracy (Eqs. (20) and (21)) with the fairness metric (either Eq. (22) or

(24)) as the final loss function of GRAPE training. Specifically, for GRAPE with resource fairness (GRAPE + $\mathcal{R}$), the final loss function is formally given by

$$\mathcal{L}_{\mathcal{R}} = \lambda_{\text{MSE}} \mathcal{L}_{\text{MSE}} + \lambda_{\text{tra}} \mathcal{L}_{\text{tra}} + \lambda_{\text{res}} \mathcal{L}_{\text{res}}, \quad (25)$$

and for GRAPE with performance fairness (GRAPE + $\mathcal{P}$), the final loss function is given by

$$\mathcal{L}_{\mathcal{P}} = \lambda_{\text{MSE}} \mathcal{L}_{\text{MSE}} + \lambda_{\text{tra}} \mathcal{L}_{\text{tra}} + \lambda_{\text{per}} \mathcal{L}_{\text{per}}, \quad (26)$$

where $(\lambda_{\text{MSE}}, \lambda_{\text{tra}}, \lambda_{\text{res}}, \lambda_{\text{per}})$ are the weights of different loss perspectives. By minimizing Eq. (25) or Eq. (26) in addition to $\mathcal{L}_{\text{MSE}}$ and $\mathcal{L}_{\text{tra}}$, we are able to regularize the model to maintain the system fairness in resource or performance. In practice, the city planner or the bike service provider can select the socioeconomic attribute as well as the fairness metric to customize for specific urban planning purposes or scenarios.

6 EXPERIMENTAL EVALUATION

We first present the evaluation setup in Sec. 6.1, followed by the experimental results in Sec. 6.2.

6.1 Evaluation Setup

• **Schemes Compared.** GRAPE is compared against the following baselines and state-of-the-art spatio-temporal approaches:

- 1) HA (Historical Average) [34]: We calculate the future bike usage based on the average of all historical data at the same time interval of a day.
- 2) ARIMA (Auto Regressive Integrated Moving Average): regresses and predicts the bike usage at the stations.
- 3) RNN (Recurrent Neural Network [14]), LSTM (Long Short-Term Memory [35]), and GRU (Gated Recurrent Units [36]): We implement these classic deep learning models for the bike usage time-series prediction.
- 4) CNN-LSTM/CNN-GRU/CNN-RNN: With the spatial bike usage in the heatmap representations, convolutional neural network (CNN) is respectively combined with LSTM/GRU/RNN to predict future bike usage.
- 5) TPA-LSTM (Temporal Pattern Attention LSTM [37]: leverages the correlations between LSTM modules for learning multivariate time series.
- 6) LSTNet (Long- and Short-term Time-series Network [38]): combines CNN and RNN to extract short- and long-term temporal patterns to predict multivariate time series, and incorporates an autoregressive model to tackle the scale insensitivity problem.
- 7) DeepST: We divide the city maps into grid maps. By using the aggregation of bike usage in each grid, we leverage spatio-temporal ResNet [39] followed by a fully connected layer to predict bike traffic at the station level.
- 8) GCNN-DDGF (Graph Convolutional Neural Network with Data-driven Graph Filter [1]): considers the adjacency matrix as an adjustable variable.

- 9) GCN-LSTM (Graph Convolutional Network (GCN) combined with LSTM for spatial and temporal bike usage characterization): The parameters of GCN and LSTM are set exactly the same as GRAPE.
- 10) GAAN [40]: predicts the bike usage with a graph neural network with an adjacency matrix attention mechanism.
- 11) GraphWaveNet [41]: contains an adaptive dependency matrix for spatial learning and a stacked dilated 1D convolution component to conduct the temporal learning.
- 12) FairST [25]: leverages the three-dimension bike usage heatmaps, two-dimension spatial information, and one-dimension external factors for equity-aware predictions with consideration of resource fairness.
- 13) Ada-STNet [42]: obtains an optimal graph adjacency matrix from both macro and micro perspectives and learns spatial relationships and temporal dependencies via a spatio-temporal convolution architecture.
- 14) CrossFormer [43]: utilizes the dependency across the bike stations for predicting the usage based on the dimension-segment-wise embeddings and the two-stage attention layers.
- 15) iTransformer [44]: leverages the transformer blocks for the bike usage prediction.
- 16) Informer [45]: adopts the efficient transformer designs for long time-series prediction.
- 17) Autoformer [46]: uses the decomposition transformers based on auto-correlations.
- 18) FEDFormer [47]: predicts the time series with the frequency enhanced decomposed transformer.
- 19) Reformer [48]: leverages an efficient transformer for the time series prediction (bike usage in our case).

• **Evaluation Metrics.** We evaluate the model prediction accuracy using MSE (Eq. (20)) and Coefficient of Determination (R^2). A higher R^2 implies a better performance [24].

We leverage Eq. (24) to evaluate the model performance fairness. To evaluate the resource fairness, we adjust Eq. (22) with the standardized prediction results without dividing by the ground-truth values, i.e.,

$$\mathcal{L}'_{\text{res}} = \frac{1}{\mathcal{F}} \sum_{k=t+1}^{t+F} \left| \frac{\sum_{i^+}^{N^+} \hat{\mathbf{z}}_{\text{tar},k}^{(d)'}[i^+]}{\sum_{i^+}^{N^+} D_{i^+}} - \frac{\sum_{i^-}^{N^-} \hat{\mathbf{z}}_{\text{tar},k}^{(d)'}[i^-]}{\sum_{i^-}^{N^-} D_{i^-}} \right|, \quad (27)$$

where we standardize the predictions $\hat{\mathbf{z}}_{\text{tar},k}^{(d)}[i]$ to get $\hat{\mathbf{z}}_{\text{tar},k}^{(d)'}[i]$ by the mean and the standard deviation of the prediction for the entire testing set. Note that if the prediction $\hat{\mathbf{z}}_{\text{tar},k}^{(d)}[i]$ significantly underestimates the ground-truth, we will get a small $\mathcal{L}'_{\text{res}}$, but it does not reflect the decrease of the resource unfairness. Therefore, by standardization we can reduce the impact of prediction accuracy on the fairness metric. The decrease in Eqs. (24) and (27) indicates a reduction of unfairness of the bike sharing system.

• **Evaluation Settings.** We evaluate GRAPE and the related schemes with the datasets presented in Sec. 2.1. We take into account the taxi systems as an auxiliary system. For all the schemes evaluated, we leverage the historical data of the past 24 hours to predict that of the next following

hour. Our bike sharing studies focus on a total of 454 stations in NYC that have bike usage every day and have not been relocated or removed throughout the year of 2019, and a total of 417 stations in Chicago that have bike usage every week and have not been relocated or removed during 2019/01 – 2019/09. We consider 50% of the user trip data for the model training and the rest half for the model testing. To construct the auxiliary graph nodes of the taxi systems, we select those regions that have their bounding boxes overlap with a 500m×500m grid area around each bike station. We note that the zones or areas are processed by the city data portal for the local planning and the analytics purposes, and the taxi pick-ups and drop-offs from the non-fixed regions are not available. Despite this, our approach and framework design is general enough to be extended to the actual GPS coordinates when the relevant data becomes available.

Our model is built and evaluated on TensorFlow 2.4.0 with CUDA 11.1. All experimental evaluations are conducted upon a server with Ryzen Threadripper 3960X 24-Core CPU and four Nvidia GeForce RTX3090 24GB GDDR6X GPUs. It takes about 20min and 10min to train GRAPE for NYC and Chicago, respectively.

• **Parameter Settings.** Unless otherwise stated, we use the following parameters by default. The embedding size of output embeddings (Eq. (8)) at each hierarchical level is set to $f_{\text{tar}}^{(p)} = f_{\text{aux}}^{(p)} = 1$. We use two-level hierarchical graphs for representation learning of both bike graph and taxi graph, i.e., $P = 2$, and set the number of nodes at the second level to be 7 for NYC and 4 for Chicago, respectively. For NYC, we use 3×1 kernel filters with 3 channels for the convolutional fusion component (Sec. 4) where 3×1 denotes 3 time intervals and 1 node. For Chicago, we use 2×1 kernel filters with 2 channels. For the spatial attention mechanism (Sec. 4), we set the channel size of the Conv1D as 516 for NYC and 254 for Chicago.

We empirically set the coefficients $\lambda_{\text{MSE}} = 1$ and $\lambda_{\text{tra}} = 0.01$ for the two prediction accuracy metrics in Eqs. (20) and (21). As Eqs. (22) and (23) have different magnitudes, we adapt λ_{res} and λ_{per} in order to obtain the similar scales in Eqs. (25) and (26). For resource fairness, we set $\lambda_{\text{res}} = 0.015$ for NYC and $\lambda_{\text{res}} = 0.001$ for Chicago regarding all three socioeconomic attributes. For performance fairness with respect to social ethnicity, income, and education, λ_{per} is (0.015, 0.018, 0.018) in NYC, and (0.075, 0.075, 0.16) in Chicago. We set the size of hidden states of LSTM, h , as 64 for NYC and 32 for Chicago. GRAPE is trained with a learning rate of 0.005 and a batch size of 64 by the Adam optimizer for 5,000 iterations.

6.2 Experimental Results

• **Overall Model Performance.** We first demonstrate the overall prediction performance in the two cities in Table 1. We can observe that GRAPE substantially outperforms the baseline models in the prediction accuracy. Existing approaches like HA, RNN, GRU, and LSTM can only capture temporal characteristics of bike usage. CNN only encodes the bike usage sequence rather than geographic information, and hence CNN and its combinations with the RNN, GRU, and LSTM may not fully capture the spatial correlations across stations and thus perform poorly in our settings.

TABLE 1: Performance on NYC and Chicago Datasets.

Schemes	NYC		Chicago	
	MSE	R^2	MSE	R^2
ARIMA	18.387	0.518	10.103	0.239
HA	17.866	0.533	11.547	0.051
RNN	12.847	0.664	5.741	0.606
GRU	12.380	0.676	5.726	0.607
LSTM	12.135	0.682	5.627	0.614
CNN-RNN	9.324	0.756	5.230	0.641
CNN-LSTM	10.503	0.725	5.646	0.613
CNN-GRU	9.968	0.739	6.491	0.555
TPA-LSTM	10.334	0.729	6.586	0.548
GCNN-DDGF	10.261	0.729	5.238	0.638
GCN-LSTM	9.857	0.742	4.775	0.673
DeepST	9.478	0.752	5.114	0.649
LSTNet	9.102	0.762	5.259	0.642
GAAN	8.877	0.769	4.973	0.659
GraphWaveNet	9.456	0.752	4.838	0.668
FairST	9.787	0.744	4.822	0.669
Ada-STNet	8.000	0.790	5.000	0.656
CrossFormer	12.294	0.637	4.542	0.539
iTransformer	10.924	0.700	4.353	0.612
Informer	9.771	0.790	4.694	0.678
Autoformer	10.389	0.777	4.198	0.712
FEDFormer	10.266	0.779	4.128	0.716
Reformer	10.410	0.776	4.133	0.716
GRAPE w/o fairness	7.898	0.793	4.090	0.720
GRAPE + $\mathcal{R}_{\text{social ethnicity}}$	8.121	0.787	4.286	0.706
GRAPE + $\mathcal{P}_{\text{social ethnicity}}$	8.214	0.785	4.477	0.693
GRAPE + $\mathcal{R}_{\text{income}}$	8.110	0.788	4.272	0.707
GRAPE + $\mathcal{P}_{\text{income}}$	8.229	0.784	4.388	0.699
GRAPE + $\mathcal{R}_{\text{education}}$	8.464	0.778	4.422	0.697
GRAPE + $\mathcal{P}_{\text{education}}$	8.406	0.780	4.524	0.690
GRAPE + $\mathcal{P}_{\text{all}}$	8.272	0.783	4.547	0.688
GRAPE + $\mathcal{R}_{\text{all}}$	8.184	0.786	4.349	0.702

We have also demonstrated that our proposed GRAPE outperformed the other state-of-the-art spatio-temporal prediction approaches. In particular, while TPA-LSTM, GCNN-DDGF, GCN-LSTM, DeepST, LSTNet, GAAN, GraphWaveNet, FairST, Ada-STNet, and other transformer-based approaches (such as CrossFormer and iTransformer) capture the spatio-temporal characteristics, they formulate the bike usage prediction as a single-modality problem. In addition, while FairST takes into account the system equity, it may underestimate the station correlations as it leverages the heatmaps as inputs, each entry of which is aggregation of total bike usage in a region. Therefore, FairST yields the lower accuracy for station-based bike usage predictions than our proposed approach.

GRAPE captures the spatial and temporal characteristics via the combination of GCN and LSTM, while the taxi system as an auxiliary graph provides additional spatial and temporal knowledge for the target graph. Such comprehensive spatio-temporal knowledge introduced by the auxiliary graph improves the accuracy of mobility prediction for the target (bike) system in our studies.

In addition to each individual attribute, i.e., social ethnicity, income, and education, we have also shown the performance combining all the three socioeconomic attributes, denoted as “all”, in Eq. (25) or Eq. (26). The accuracy of our GRAPE with fairness regularization, denoted as “GRAPE w/o fairness”, outperforms all the baselines, despite the negligible accuracy degradation compared with the one without regularization, i.e., on average 2.81% lower in NYC and 5.49% lower in Chicago.

The performance of the models varies across the two cities, likely due to their large difference in the total number of bike usage and resultant bike mobility patterns as discussed in Sec. 2.1. Despite this variation, GRAPE can still

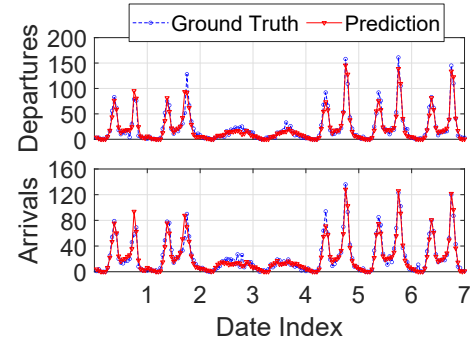


Fig. 11: Hourly bike-usage (departures and arrivals) prediction from July 8th to July 14th for a station at Manhattan, NYC.

TABLE 2: Values of $\mathcal{L}'_{\text{res}}$ with 3 socioeconomic attributes.

Attributes	Models	NYC	Chicago
Social Ethnicity	w/o $\mathcal{R}$	311.802	190.260
	w/ $\mathcal{R}$	142.788	128.630
Income	w/o $\mathcal{R}$	302.246	212.045
	w/ $\mathcal{R}$	229.720	163.882
Education	w/o $\mathcal{R}$	425.122	291.925
	w/ $\mathcal{R}$	343.866	238.784

TABLE 3: Values of $\mathcal{L}_{\text{per}}$ with 3 socioeconomic attributes.

Attributes	Models	NYC	Chicago
Social Ethnicity	w/o $\mathcal{P}$	143.232	22.917
	w/ $\mathcal{P}$	137.039	17.758
Income	w/o $\mathcal{P}$	153.545	47.608
	w/ $\mathcal{P}$	134.756	39.597
Education	w/o $\mathcal{P}$	172.491	53.047
	w/ $\mathcal{P}$	131.926	47.621

achieve excellent accuracy for both cities. Fig. 11 also shows GRAPE’s accurate hourly bike-usage predictions at a station at Manhattan, NYC during a week.

• **Fairness Evaluation.** We further evaluate fairness in resources and model performance by Eqs. (27) and (24) regarding each of the three socioeconomic attributes considered. This way, we can specifically evaluate the interaction of the model with each socioeconomic attribute. The results are shown in Tables 2 and 3. GRAPE with resource fairness (“w/ $\mathcal{R}$ ”) and the GRAPE with performance fairness (“w/ $\mathcal{P}$ ”) are compared with GRAPE without consideration of fairness (denoted as “w/o $\mathcal{R}$ ” and “w/o $\mathcal{P}$ ”, respectively). Combined with Table 1, we can see that GRAPE reduces the inequity in resources and model performance of the systems while maintaining the prediction accuracy for all the three socioeconomic attributes in NYC and Chicago.

The difference of the same fairness metric between the attributes may result from the diverse socioeconomic characteristics of bike stations. For example, as we recall from Fig. 7, some stations are overlapped by disadvantaged communities in terms of education and the advantaged communities in terms of income. Comparing the fairness performance of GRAPE between the two cities, we find that all metrics of Chicago are smaller than that of NYC. This is likely due to the overall much lower bike usage in Chicago than in NYC. With lower bike usage, we see that the (un)fairness metrics of resources (based on Eqs. (23) and (27)) as well as the performance in Chicago are overall smaller than in NYC.

We further compare the fairness metrics, $\mathcal{L}'_{\text{res}}$ and

$\mathcal{L}_{\text{per}}$, of GRAPE with the four state-of-the-art baselines, GraphWaveNet, DeepST, GAAN, and LSTNet, for social ethnicity in NYC as an example which is illustrated in Fig. 12. We can see that GRAPE obtains a significant improvement of equity of social ethnicity in the city compared to other traditional models which fail to consider system fairness and, therefore, generate biased predictions. The results demonstrate the effectiveness of GRAPE's equity-aware designs.

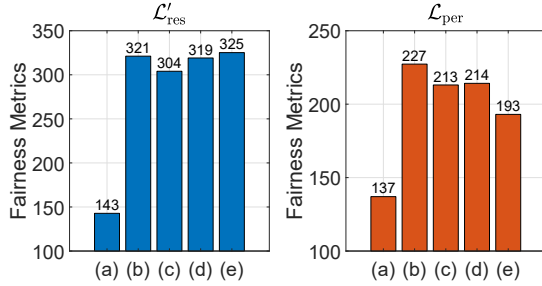


Fig. 12: $\mathcal{L}'_{\text{res}}$ and $\mathcal{L}_{\text{per}}$ w.r.t. social ethnicity in NYC for (a) GRAPE, (b) GraphWaveNet, (c) DeepST, (d) GAAN, and (e) LSTNet.

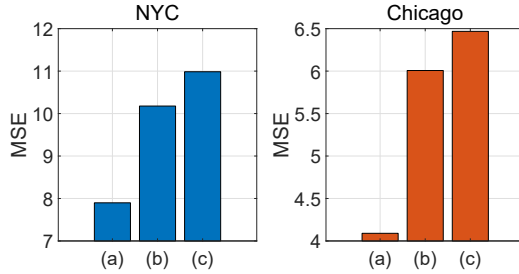


Fig. 13: Ablation study of GRAPE.

• **Ablation Study.** We have further conducted ablation studies by comparing GRAPE with complete modules (denoted as (a)) with the following two major variations:

- Without temporal and spatial learning (denoted as (b)): We omit the temporal and spatial learning module while keeping other parts of the model the same as GRAPE.
- Without auxiliary system learning (denoted as (c)): We remove the fusion of auxiliary (taxi) system inputs based on graph differentiable pooling and convolutional fusion.

We evaluate these variants of GRAPE for the datasets of NYC and Chicago without considering fairness regularization. Fig. 13 shows the MSEs of the studied schemes. From the figure we observe a better performance of GRAPE than the variants, demonstrating the effectiveness of the temporal and spatial learning and the auxiliary system.

• **Sensitivity Studies.** We first demonstrate in Table 4 the performance of all the schemes during weekdays and weekends without considering fairness for one week's data of Chicago. We note that the negative R^2 of weekend prediction by ARIMA indicates its prediction does not reflect the trend of the bike-usage patterns. We can observe that GRAPE outperforms all baselines for usage predictions of both weekdays and weekends, demonstrating the effectiveness of our model designs in handling both weekday and weekend dynamics. We also observe overall lower errors

TABLE 4: Weekday/Weekend performance (Chicago).

Schemes	Weekdays		Weekends	
	MSE	R^2	MSE	R^2
ARIMA	11.769	0.234	15.214	-0.008
HA	7.452	0.330	14.581	0.034
RNN	4.795	0.605	5.729	0.620
GRU	4.599	0.621	6.420	0.575
LSTM	4.746	0.609	6.027	0.601
CNN-RNN	4.633	0.618	4.689	0.689
CNN-LSTM	4.633	0.618	6.839	0.547
CNN-GRU	5.728	0.528	5.859	0.612
TPA-LSTM	4.888	0.597	7.736	0.487
GCNN-DDGF	4.880	0.598	5.862	0.612
GCN-LSTM	3.822	0.685	4.834	0.680
DeepST	5.081	0.574	5.269	0.693
LSTNet	4.899	0.677	7.000	0.536
GAAN	4.012	0.669	5.095	0.662
GraphWaveNet	4.670	0.615	5.113	0.661
FairST	3.968	0.673	5.585	0.630
GRAPE	3.614	0.702	4.602	0.695

for weekdays than weekends for all the schemes. This is due mainly to the potentially higher complexity of human mobility patterns during weekends (e.g., people may have various travel or entertainment activities) than on weekdays (e.g., major activities related to commute routines).

Fig. 14a further illustrates the impact of the length of future time intervals $\mathcal{F}$ upon GRAPE (without fairness metrics). From the figure, we can see that as the length of predicted intervals increases, the overall prediction accuracy decreases. This is because the correlations between the target intervals in a far prediction horizon and the past ones in the input are more complex than those between the intervals in a near future. GRAPE can still achieve reasonable accuracy for the far prediction horizon.

Using NYC as an example, we have conducted a sensitivity study on MSE, $\mathcal{L}'_{\text{res}}$ and $\mathcal{L}_{\text{per}}$ with respect to their coefficients in Eqs. (25) and (26). For resource fairness, we study Chicago's social ethnicity and vary λ_{res} to be 0.0005, 0.001, 0.0015, 0.002, and 0.0025, with a default setting of λ_{per} as 0.015 (Sec. 6.1). For performance fairness, we study NYC's social ethnicity and vary λ_{per} to be 0.01, 0.015, 0.02, 0.025, and 0.03, with a default setting of λ_{res} as 0.015 (Sec. 6.1). Figs. 14b and 14c show that increasing the weight of resource or performance fairness loss reduces the corresponding metric, with a slight increase in the prediction error. For the performance fairness, we can observe an increase of the metric as the coefficient increases, which is likely due to large errors in model performance. Considering both prediction accuracy and fairness metrics, we pick 0.001 for λ_{res} and 0.015 for λ_{per} by default to maintain the accuracy of prediction while improving the fairness.

• **Result Visualization and Case Studies.** We have conducted the following studies to further understand the model behaviors of GRAPE.

(a) **Node Assignment.** We visualize in Fig. 15 the transition matrix at the input level, i.e., $\mathbf{T}_{\text{tar},k}^{(1)}$ in Eq. (7), to show the probability values of assigning bike stations to two hierarchical nodes. The darker color represents the higher probability values that an input graph node to a node in higher hierarchy. We can see that bike stations across NYC

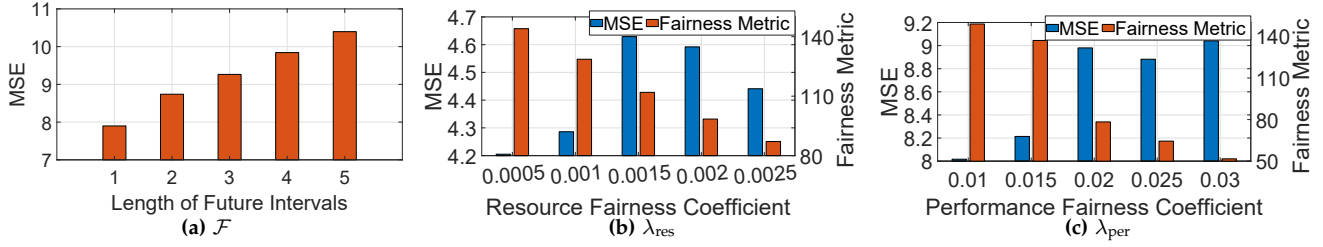


Fig. 14: The sensitivity studies of GRAPE: (a) length of future time intervals; (b) MSE and fairness vs. λ_{res} ; and (c) MSE and fairness vs. λ_{per} .

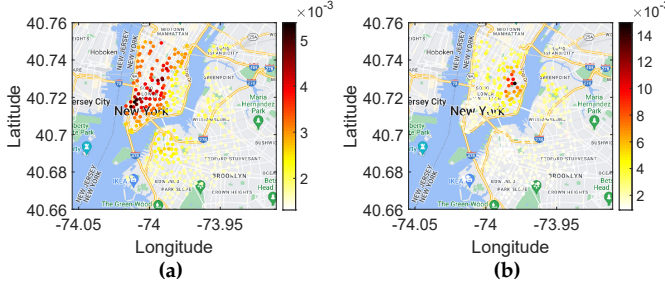


Fig. 15: Probability distribution of bike station assignment to two different hierarchical nodes, i.e., the transition matrix at the input level, $T_{tar,k}^{(1)}$ (Eq. 7), on 12pm, 2019/09/01.

are grouped together into different regions, each of which is represented by a hierarchical node. Such a node assignment behavior help GRAPE capture the shared spatio-temporal correlations among the bike sharing station nodes, yielding

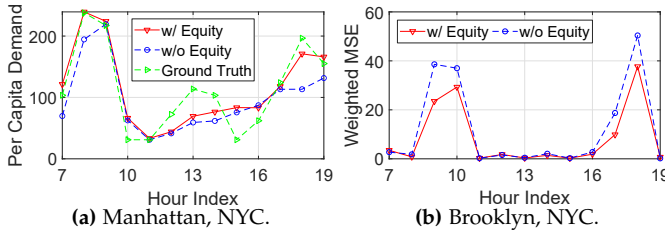


Fig. 16: The interaction comparison of (a) forecasted *per capita* demand (along with ground-truth) and (b) the prediction MSE weighted by population as a function of time with and without considering (a) resource fairness and (b) performance fairness for two stations on 2019/07/12 (Friday) within the educational disadvantaged communities in NYC.

(b) *Equity-Awareness*. Taking a station within an educational disadvantaged community in Manhattan, NYC as an example, Fig. 16a illustrates its hourly *per capita* demand, $\frac{\hat{z}_{tar,k}^{(d)}[i^-]}{\sum_{i^-} D_{i^-}}$, over 12 hours of a day. We compare the *per capita* demand calculated by GRAPE (in red solid lines) with that of GRAPE without fairness consideration (in blue dashed lines). Recall that the *per capita* demand of a station is used in Eq. (27) to calculate the fairness metric. With the regularized predictions, one may expect that more bike resources can be allocated to these stations within disadvantaged communities (as illustrated in Fig. 7b) via the equity-aware city planning. Fig. 16a further shows a higher *per capita* demand with the regularized predictions, corroborating the effectiveness of GRAPE in improving the equity of resource distribution. In addition, we can also observe that the predicted *per capita* demand of GRAPE (in red solid lines) captures the

overall trends of the ground-truth (in green dashed lines), demonstrating the effectiveness of GRAPE in learning the *per capita* demands.

Similarly, Fig. 16b illustrates MSEs weighted by *population* over 12 hours of a day for a bike station near the disadvantaged communities in terms of education around Brooklyn, NYC, i.e., $\frac{D_{i^-}}{2\mathbb{E}[D_{i^-}]} \sum_{q \in d,r} (\hat{z}_{tar,k}^{(q)}[i^-] - z_{tar,k}^{(q)}[i^-])^2$. One can observe that the lower weighted MSEs of those stations in the disadvantaged community, demonstrating the effectiveness of GRAPE in reducing the model performance unfairness.

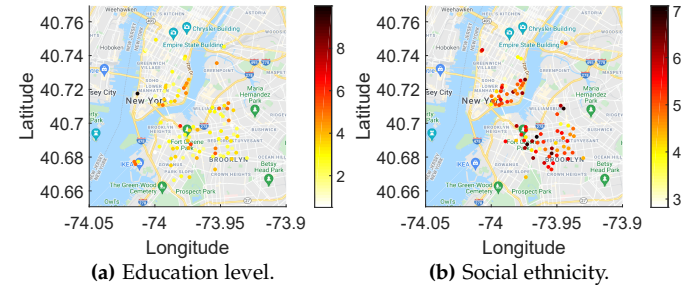


Fig. 17: Station-level resource unfairness reduction for stations around disadvantaged communities in NYC. Each dot represents a station, and warmer colors imply more reduction.

We finally visualize the reduction of resource unfairness in Fig. 17 at those bike stations around the disadvantaged communities in terms of education level (a) and social ethnicity (b) in NYC. The warmer colors indicate a larger reduction in the unfairness. We can observe the unfairness reduction by introducing our fairness regularization, demonstrating the spatial resource fairness improvements in different communities.

7 DISCUSSION

• **Integrating Other Mobility Modalities.** In the current prototype study, we only consider taxi as the auxiliary systems and bike sharing as the target system for predicting bike usage due to the lack of data of other transportation platforms. As the transportation system of a city contains many other platforms such as ride sharing and public transits like bus and subway, other transportation systems may also provide useful spatio-temporal information to the target system and benefit the prediction tasks as well as the entire city planning. The framework designs in GRAPE can easily incorporate other types of transportation platforms [49]–[53] into the target system for mobility prediction tasks when available. Many cities like NYC and Chicago

are embracing the open data initiatives to coordinate release of multi-source mobility data (e.g., bike sharing and taxis). Our research here can provide new insights when integrated with these emerging efforts for greater good of the urban transportation network.

• Inclusion of Additional Socioeconomic Attributes.

While our study focused on the three socioeconomic attributes of social ethnicity, income, and education, our model is general enough to be extended to other important factors such as age and gender. In future, we would like to further investigate the impacts of multiple socioeconomic attributes for broader application scenarios.

8 RELATED WORK

• **Bike Usage Prediction.** Bike-usage prediction is essential for the bike sharing system management [17], [18], [54]. However, the station-level prediction is challenging due to the complex spatio-temporal characteristics at a large number of stations in a city. Prior studies [1]–[4], [40], [55]–[57] did not consider incorporation of the mobility of other modalities of transportation. While information retrieval and time series analysis based on multiple modalities have been explored [58]–[65] through various learning paradigms, how to realize the fine-grained station-based bike usage prediction, particularly through a paradigm of fusing the constructed graphs, remains largely under-explored.

Different from these prior studies, our work here focuses on a novel cross-modality graph fusion design that leverages the spatio-temporal correlations between taxi and bike sharing networks and constructs the auxiliary graph feature to enhance the target bike sharing prediction at the finer granularity, i.e., station level or hourly basis. Specifically, we have designed a novel cross-modality graph fusion mechanism with the target and auxiliary graphs, which combines the graphs at the hierarchical levels, extracts their mobility patterns by graph differentiable pooling [28], and thus yields the high usage prediction accuracy.

• **Mobility Fairness.** Besides operational efficiency, the fairness of the mobility systems such as bike sharing has also become an important factor that many city planners must consider. Like other applications such as classification, fairness in machine learning has attracted significant attention recently [66]–[75], where the models were usually developed in a multi-task fashion to maintain the utility while preventing discrimination [67] regarding sensitive socioeconomic attributes. Specifically, in spatio-temporal settings such as dockless bike-usage prediction [25], [33], [76] and e-scooter sharing [5] fairness is considered as the closeness of the prediction results for the advantaged and disadvantaged communities, and the variance in demand across the two groups is considered to be the effect of social inequity.

However, these prior studies did not consider *differentiating* the regional disparities in model performance based on the population ratio to mitigate the potential biases towards populated regions. Furthermore, how the fairness metrics *interact* with the spatial and temporal dynamics of mobility predictions, particularly in the station-based bike sharing systems, remain largely unexplored [12], [13]. Our studies here further fill the above-mentioned research gaps

by integrating the regional disparity with differentiation of population ratio, as well as in-depth spatial and temporal interaction analysis. Our experimental studies (Sec. 6) have further corroborated the effectiveness and accuracy of our proposed designs.

9 CONCLUSION

In this paper, we have proposed GRAPE, an equity-aware graph-fusion differentiable pooling network, to jointly predict station-level bike usage and improve the bike sharing system social equity. We have incorporated the taxi network as an auxiliary system to the target bike sharing network through a graph-fusion differentiable pooling algorithm, where two hierarchical graph networks extract the spatial characteristics of the target and the auxiliary graphs, and the extracted embeddings are fused via a novel convolutional fusion design. The temporal characteristics of the fused embeddings are further captured by an LSTM network. To mitigate the unfairness of the bike sharing networks, we have proposed fairness metrics in terms of resources and performance. Extensive experiments on the real-world datasets from NYC and Chicago have shown that GRAPE predicts the bike usage with high accuracy and substantially improves the resource and performance fairness.

ACKNOWLEDGMENT

We would like to acknowledge Hao Wang, Yuyao Li, and Yawen Deng for their contributions to the additional experimental studies. This project is supported, in part, by the National Science Foundation (NSF) under Grant 2303575, the Google Research Scholar Award Program (2021–2022), the NVIDIA Applied Research Accelerator Program Award (2021–2022), and the University of Connecticut (UConn) Research Excellence Program Awards (2020–2021, 2022–2024). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES

- [1] 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," *Transportation Research Part C: Emerging Technologies*, vol. 97, pp. 258–276, 2018.
- [2] Y. Li and Y. Zheng, "Citywide bike usage prediction in a bike-sharing system," *IEEE TKDE*, vol. 32, no. 6, pp. 1079–1091, 2019.
- [3] S. He and K. G. Shin, "Towards fine-grained flow forecasting: A graph attention approach for bike sharing systems," in *Proc. WWW*, 2020, pp. 88–98.
- [4] H. Liu, Q. Wu, F. Zhuang, X. Lu, D. Dou, and H. Xiong, "Community-aware multi-task transportation demand prediction," in *Proc. AAAI*, vol. 35, no. 1, 2021, pp. 320–327.
- [5] S. He and K. G. Shin, "Socially-equitable interactive graph information fusion-based prediction for urban dockless e-scooter sharing," in *Proc. WWW*, 2022, p. 3269–3279.
- [6] C. Shi, X. Kong, Y. Huang, S. Y. Philip, and B. Wu, "HeteSim: A general framework for relevance measure in heterogeneous networks," *IEEE TKDE*, vol. 26, no. 10, pp. 2479–2492, 2014.
- [7] W. Zhang, Y. Fang, Z. Liu, M. Wu, and X. Zhang, "Mg2Vec: learning relationship-preserving heterogeneous graph representations via metagraph embedding," *IEEE TKDE*, 2020.
- [8] Y. Yang, Z. Guan, J. Li, W. Zhao, J. Cui, and Q. Wang, "Interpretable and efficient heterogeneous graph convolutional network," *IEEE TKDE*, 2021.

- [9] Z. Yang, J. Hu, Y. Shu, P. Cheng, J. Chen, and T. Moscibroda, "Mobility modeling and prediction in bike-sharing systems," in *Proc. ACM MobiSys*, 2016, pp. 165–178.
- [10] L. Chen, D. Zhang, L. Wang, D. Yang, X. Ma, S. Li, Z. Wu, G. Pan, T.-M.-T. Nguyen, and J. Jakubowicz, "Dynamic cluster-based over-demand prediction in bike sharing systems," in *Proc. ACM UbiComp*, 2016, pp. 841–852.
- [11] D. Duran-Rodas, D. Villeneuve, and G. Wulforst, "Bike-sharing: the good, the bad, and the future: an analysis of the public discussion on twitter," *European Journal of Transport and Infrastructure Research*, vol. 20, no. 4, pp. 38–58, 2020.
- [12] J. Ursaki, L. Aultman-Hall *et al.*, "Quantifying the equity of bike-share access in us cities," University of Vermont. Transportation Research Center, Tech. Rep., 2015.
- [13] K. Hosford and M. Winters, "Who are public bicycle share programs serving? An evaluation of the equity of spatial access to bicycle share service areas in Canadian cities," *Transportation Research Record*, vol. 2672, no. 36, pp. 42–50, 2018.
- [14] Y. Pan, R. C. Zheng, J. Zhang, and X. Yao, "Predicting bike sharing demand using recurrent neural networks," *Procedia Computer Science*, vol. 147, pp. 562–566, 2019.
- [15] R. W. Aldridge, D. Lewer, S. V. Katikireddi, R. Mathur, N. Pathak, R. Burns, E. B. Fragaszy, A. M. Johnson, D. Devakumar, I. Abubakar *et al.*, "Black, Asian and minority ethnic groups in england are at increased risk of death from COVID-19: indirect standardisation of NHS mortality data," *Wellcome Open Research*, vol. 5, no. 88, p. 88, 2020.
- [16] W. N. L. Pirtle, "Racial capitalism: A fundamental cause of novel coronavirus (COVID-19) pandemic inequities in the united states," *Health Education & Behavior*, 2020.
- [17] X. Yang, S. He, and M. Tabatabaie, "Equity-aware cross-graph interactive reinforcement learning for bike station network expansion," in *Proc. ACM SIGSPATIAL*, 2023.
- [18] S. He and K. G. Shin, "Information fusion for (re) configuring bike station networks with crowdsourcing," *IEEE TKDE*, 2020.
- [19] *Citi Bike in NYC*, 2021. [Online]. Available: <https://www.citibikenyc.com>
- [20] *Divvy Bike in Chicago*, 2021. [Online]. Available: <https://www.divvybikes.com>
- [21] *NYC Taxis*, 2021. [Online]. Available: <https://www1.nyc.gov/site/tlc/index.page>
- [22] *Chicago Taxis*, 2021. [Online]. Available: <https://data.cityofchicago.org>
- [23] *US Census Tract*, 2021. [Online]. Available: <https://www.census.gov/>
- [24] N. R. Draper and H. Smith, *Applied regression analysis*. John Wiley & Sons, 1998, vol. 326.
- [25] A. Yan and B. Howe, "FairST: Equitable spatial and temporal demand prediction for new mobility systems," in *Proc. ACM SIGSPATIAL*, 2019, pp. 552–555.
- [26] Y. Xu and P. Xu, "Trading the system efficiency for the income equality of drivers in rideshare," *arXiv preprint arXiv:2012.06850*, 2020.
- [27] D. Chai, L. Wang, and Q. Yang, "Bike flow prediction with multi-graph convolutional networks," in *Proc. ACM SIGSPATIAL*, 2018, pp. 397–400.
- [28] Z. Ying, J. You, C. Morris, X. Ren, W. Hamilton, and J. Leskovec, "Hierarchical graph representation learning with differentiable pooling," in *Proc. NeurIPS*, 2018, pp. 4800–4810.
- [29] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.
- [30] X. Yang, S. He, B. Wang, and M. Tabatabaie, "Spatio-Temporal Graph Attention Embedding for Joint Crowd Flow and Transition Predictions: A Wi-Fi-based Mobility Case Study," *Proc. ACM IMWUT*, vol. 5, no. 4, dec 2022.
- [31] N. Wu, X. W. Zhao, J. Wang, and D. Pan, "Learning effective road network representation with hierarchical graph neural networks," in *Proc. ACM SIGKDD*, 2020, pp. 6–14.
- [32] C. F. Van Loan and G. Golub, "Matrix computations (Johns Hopkins studies in mathematical sciences)," 1996.
- [33] A. Yan and B. Howe, "Fairness-aware demand prediction for new mobility," in *Proc. AAAI*, vol. 34, no. 01, 2020, pp. 1079–1087.
- [34] J. E. Froehlich, J. Neumann, and N. Oliver, "Sensing and predicting the pulse of the city through shared bicycling," in *Proc. IJCAI*, 2009.
- [35] T. Lin, B. G. Horne, P. Tino, and C. L. Giles, "Learning long-term dependencies in NARX recurrent neural networks," *IEEE TNNLS*, vol. 7, no. 6, pp. 1329–1338, 1996.
- [36] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.
- [37] S.-Y. Shih, F.-K. Sun, and H.-y. Lee, "Temporal pattern attention for multivariate time series forecasting," *Machine Learning*, vol. 108, no. 8, pp. 1421–1441, 2019.
- [38] G. Lai, W.-C. Chang, Y. Yang, and H. Liu, "Modeling long-and short-term temporal patterns with deep neural networks," in *Proc. ACM SIGIR*, 2018, pp. 95–104.
- [39] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proc. AAAI*, 2017.
- [40] X. Yang, S. He, and H. Huang, "Station correlation attention learning for data-driven bike sharing system usage prediction," in *Proc. IEEE MASS*, 2020.
- [41] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, "Graph wavenet for deep spatial-temporal graph modeling," *arXiv preprint arXiv:1906.00121*, 2019.
- [42] X. Ta, Z. Liu, X. Hu, L. Yu, L. Sun, and B. Du, "Adaptive spatio-temporal graph neural network for traffic forecasting," *Knowledge-Based Systems*, vol. 242, p. 108199, 2022.
- [43] Y. Zhang and J. Yan, "CrossFormer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting," in *Proc. ICLR*, 2022.
- [44] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, "iTransformer: Inverted transformers are effective for time series forecasting," *arXiv preprint arXiv:2310.06625*, 2023.
- [45] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proc. AAAI*, vol. 35, no. 12, 2021, pp. 11 106–11 115.
- [46] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 419–22 430, 2021.
- [47] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "FedFormer: Frequency enhanced decomposed transformer for long-term series forecasting," in *Proc. ICML*. PMLR, 2022, pp. 27 268–27 286.
- [48] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," *arXiv preprint arXiv:2001.04451*, 2020.
- [49] S. Wang, T. He, D. Zhang, Y. Liu, and S. H. Son, "Towards efficient sharing: A usage balancing mechanism for bike sharing systems," in *Proc. WWW*. ACM, 2019, pp. 2011–2021.
- [50] S. He and K. G. Shin, "Dynamic flow distribution prediction for urban dockless e-scooter sharing reconfiguration," in *Proc. WWW*, 2020, pp. 133–143.
- [51] K. Ouyang, Y. Liang, Y. Liu, Z. Tong, S. Ruan, D. Rosenblum, and Y. Zheng, "Fine-grained urban flow inference," *IEEE TKDE*, 2020.
- [52] Z. Pan, W. Zhang, Y. Liang, W. Zhang, Y. Yu, J. Zhang, and Y. Zheng, "Spatio-temporal meta learning for urban traffic prediction," *IEEE TKDE*, pp. 1–1, 2020.
- [53] L. Lin, J. Li, F. Chen, J. Ye, and J. Huai, "Road traffic speed prediction: A probabilistic model fusing multi-source data," *IEEE TKDE*, vol. 30, no. 7, pp. 1310–1323, 2018.
- [54] D. A. Tedjopurnomo, Z. Bao, B. Zheng, F. Choudhury, and A. K. Qin, "A survey on modern deep neural network for traffic prediction: Trends, methods and challenges," *IEEE TKDE*, pp. 1–1, 2020.
- [55] P. Hulot, D. Aloise, and S. D. Jena, "Towards station-level demand prediction for effective rebalancing in bike-sharing systems," in *Proc. ACM SIGKDD*, 2018, pp. 378–386.
- [56] Y. Li, Z. Zhu, D. Kong, M. Xu, and Y. Zhao, "Learning heterogeneous spatial-temporal representation for bike-sharing demand prediction," in *Proc. AAAI*, vol. 33, 2019, pp. 1004–1011.
- [57] J. Sun, J. Zhang, Q. Li, X. Yi, Y. Liang, and Y. Zheng, "Predicting citywide crowd flows in irregular regions using multi-view graph convolutional networks," *IEEE TKDE*, pp. 1–1, 2020.
- [58] L. Wu, Y. Wang, and L. Shao, "Cycle-consistent deep generative hashing for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1602–1612, 2018.
- [59] D. Chen, M. Wang, H. Chen, L. Wu, J. Qin, and W. Peng, "Cross-modal retrieval with heterogeneous graph embedding," in *Proc. ACM MM*, 2022, pp. 3291–3300.

- [60] Y. Liang, G. Huang, and Z. Zhao, "Joint demand prediction for multimodal systems: A multi-task multi-relational spatiotemporal graph neural network approach," *Transportation Research Part C: Emerging Technologies*, vol. 140, p. 103731, 2022.
- [61] F. Rodrigues, I. Markou, and F. C. Pereira, "Combining time-series and textual data for taxi demand prediction in event areas: A deep learning approach," *Information Fusion*, vol. 49, pp. 120–129, 2019.
- [62] 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. ACM SIGKDD*, 2019, pp. 305–313.
- [63] J. Deng, X. Chen, Z. Fan, R. Jiang, X. Song, and I. W. Tsang, "The pulse of urban transport: exploring the co-evolving pattern for spatio-temporal forecasting," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 15, no. 6, pp. 1–25, 2021.
- [64] S. Wang, H. Miao, J. Li, and J. Cao, "Spatio-temporal knowledge transfer for urban crowd flow prediction via deep attentive adaptation networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 5, pp. 4695–4705, 2021.
- [65] 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 International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2022, pp. 857–862.
- [66] S. Hajian and J. Domingo-Ferrer, "A methodology for direct and indirect discrimination prevention in data mining," *IEEE TKDE*, vol. 25, no. 7, pp. 1445–1459, 2012.
- [67] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proc. ACM ITCS*, 2012, pp. 214–226.
- [68] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proc. ICML*, 2013, pp. 325–333.
- [69] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proc. NeurIPS*, 2016, pp. 3315–3323.
- [70] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Proc. NeurIPS*, 2016, pp. 4349–4357.
- [71] A. Khademi, S. Lee, D. Foley, and V. Honavar, "Fairness in algorithmic decision making: An excursion through the lens of causality," in *Proc. WWW*, 2019, pp. 2907–2914.
- [72] Y. Wu, L. Zhang, and X. Wu, "On convexity and bounds of fairness-aware classification," in *Proc. WWW*, 2019, pp. 3356–3362.
- [73] T. Zhang, J. Li, M. Han, W. Zhou, P. Yu *et al.*, "Fairness in semi-supervised learning: Unlabeled data help to reduce discrimination," *IEEE TKDE*, 2020.
- [74] Z. Chen, P. Cheng, L. Chen, X. Lin, and C. Shahabi, "Fair task assignment in spatial crowdsourcing," *Proc. VLDB*, vol. 13, no. 12, 2020.
- [75] Y. Zhao, K. Zheng, J. Guo, B. Yang, T. B. Pedersen, and C. S. Jensen, "Fairness-aware task assignment in spatial crowdsourcing: Game-theoretic approaches," in *Proc. IEEE ICDE*, 2021, pp. 265–276.
- [76] A. Yan and B. Howe, "FairST: Equitable Spatial and Temporal Demand Prediction for New Mobility Systems," in *Proc. ACM SIGSPATIAL*, 2019, p. 552–555.